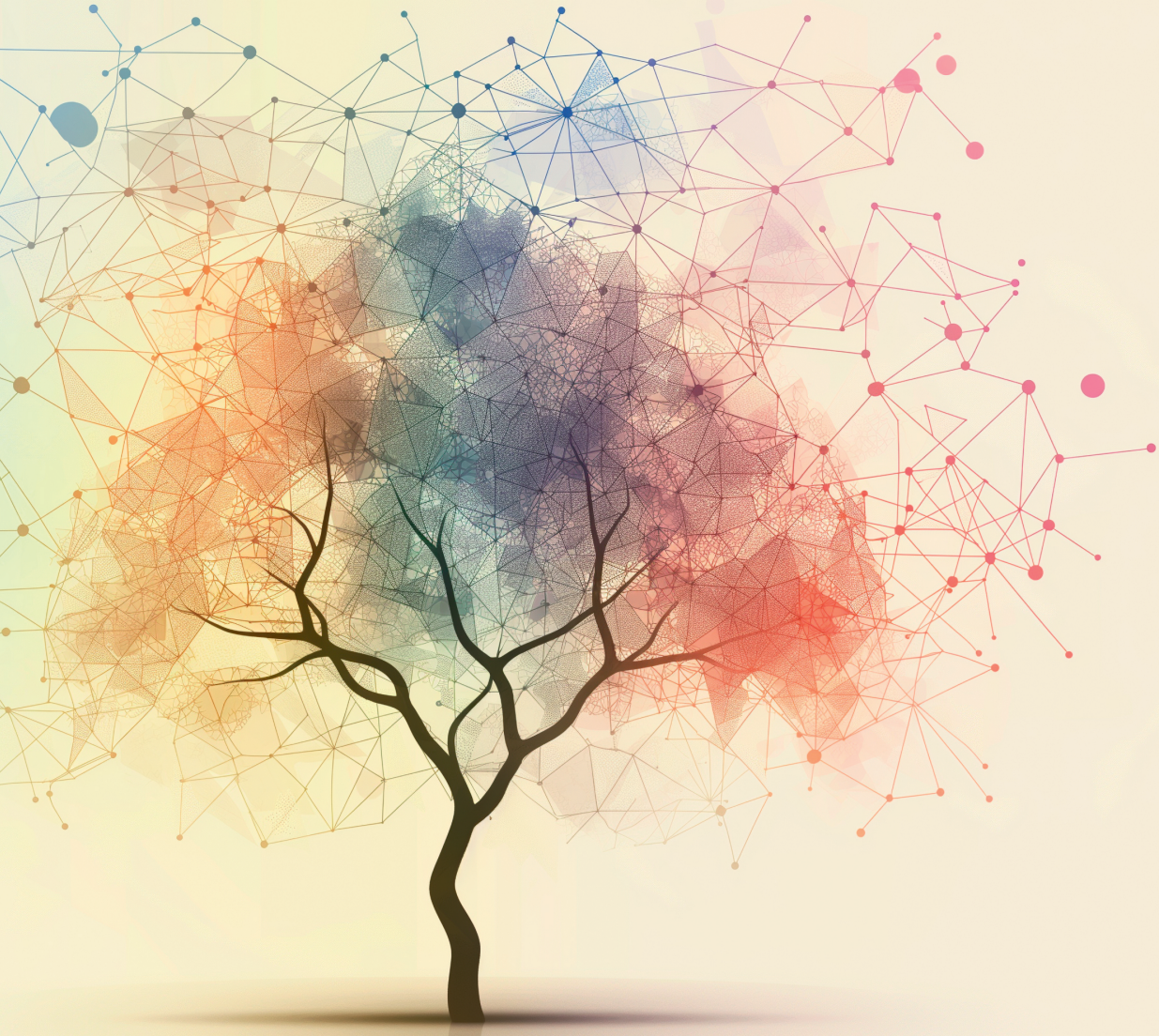
Tracing the evolutionary history of cellular life

Tara A. Mahendrarajah



Tracing the evolutionary history of cellular life

Tara A. Mahendrarajah

COLOFON

Copyright 2024 © Tara Mahendrarajah

All rights reserved. No parts of this thesis may be reproduced, stored in a retrieval system or transmitted in any form or by any means without permission of the author.

ISBN: 978-94-93391-22-2
Printed by Proefschriftspecialist | proefschriftspecialist.nl
Layout and design: Anna Bleeker | persoonlijkproefschrift.nl

Tracing the evolutionary history of cellular life

Het traceren van de evolutionaire geschiedenis van cellulair leven

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen op

vrijdag 13 september 2024 des middags te 12.15 uur

door

Tara Avanthi Mahendrarajah

geboren op 12 september 1989
te Newton, Massachusetts, Verenigde Staten

Promotoren:

Prof. dr. A. Spang
Prof. dr. L. Villanueva

Beoordelingscommissie:

Prof. dr. B.E. Dutilh
Prof. dr. T.J.G. Ettema
Dr. T. Gabaldón
Prof. L.A. Katz
Prof. dr. B. Snel

What do you see when you look at a tree? Leaves and twigs and branches?
Or do you see a living thing that moves and breathes and dances?

- Emma Carlisle

TABLE OF CONTENTS

SUMMARY

The Earth formed around 4.5 billion years ago, with life predicted to have appeared on the planet not long after. As we observe our world today, we can see it teeming with a wide variety of organisms varying greatly in size, complexity, and lifestyle. Remarkably, despite such immense diversity with the biosphere, all living things are united by a set of commonalities that hint to a single shared ancestry of all life. That is, all cellular life on Earth as we know it descended from a single primordial entity known as the last universal common ancestor (LUCA). What and who LUCA was, and how life on Earth diversified into modern organisms has long captivated scientists seeking to understand our origins.

A central approach to this profound question has focused on visualizing the known diversity of life on the planet within a schema that provides insight into evolutionary relationships. In the 1850s, Charles Darwin had speculated that all living things were related to one another through time, and that these connections can be organized in a tree-like network more commonly referred to as a tree of life (TOL). Though the TOL had meager beginnings as sketches in Darwin's notebooks, it has since become the centerpiece of phylogenetics, or the field devoted to studying the evolutionary history among and between organisms. Fundamentally, the TOL is a radiating diagram assembled from branches and nodes that represent life's evolutionary history from our microbial ancestors to complex modern life, with LUCA sitting at the root. Early manifestations of the TOL were constructed using visual observation of morphological traits, however the genetic sequencing revolution of the last century has uncovered an enormous diversity of organisms invisible to the naked eye that make up a large proportion of the biosphere. Inclusion of these lineages, many of which remain uncultivated, within the TOL have profoundly reshaped our understanding of life's evolutionary course.

Centuries later, the TOL remains a valuable framework from which to study important aspects of cellular evolution. The first molecular TOL, that is a tree inferred using the evolutionary history of gene sequences, was presented in 1990 by Carl Woese and coworkers, who demonstrated that all life on Earth belongs to one of three major domains of life, the Archaea, Bacteria, or Eukaryota. Major differences between these groups include cell morphology and complexity with Archaea and Bacteria comprising cells without a nucleus and are therefore known as prokaryotes, while Eukaryota or eukaryotes contain a nucleus and specialized organelles. This initial tripartite, or three-domain tree (3D tree), remained the central organization of life within the field of phylogenetics for many decades. Recent updates on this molecular TOL invited an alternative evolutionary scenario to the 3D tree, suggesting instead that cellular evolution is best reflected in a two-domain tree (2D tree) arrangement where the Archaea and Bacteria are primary domains descending from LUCA. In contrast, eukaryotes form a secondary domain that originated through a merger of ancestral prokaryotic cells through a process known as eukaryogenesis.

Greater accessibility to genetic information and larger datasets have necessitated improved computational methods that can handle and accurately model the data. At the broadest level, the work presented in this thesis contributes new tools and approaches used to construct the TOL and leads to knowledge that improves the interpretation of evolutionary relationships in light of newly discovered organisms. The chapters of this thesis address open and unresolved questions of life's evolutionary course from LUCA to eukaryogenesis, with a focus on the shape of the TOL and the position of LUCA, the timing of major cellular transitions, and the metabolic potential of key ancestors.

Critical and at times controversial debate over the structure of the TOL is ongoing to this day, such as the disparity between the 3D and 2D tree hypotheses and more recent discussions over the distance between the two primary domains, Archaea and Bacteria. Since the prokaryotic lineages are the first to diverge and radiate from LUCA, it is pertinent to properly resolve their relative positions in the tree. Traditional phylogenetic methods that utilized a small collection of marker genes (~50) to infer the TOL resolve a long interdomain branch between Archaea and Bacteria, suggesting substantial evolutionary change between the two. However, a thorough analysis of a variety of marker genes, including a greatly expanded set of metabolic genes, suggests that the length of this branch is sensitive to the marker genes that are selected. More specifically, genes that have complicated evolutionary histories, such as those that are frequently exchanged between Archaea and Bacteria, artificially draw the domains closer together. My analyses find that careful gene and sequence selection is required to overcome phylogenetic artifacts and that the best performing marker genes for estimating the deep relationship between Archaea and Bacteria are vertically evolving genes.

Aside from the shape of the TOL, the timing of major events in cellular evolution is poorly resolved, namely due to the paucity of prokaryotic fossils. Microorganisms, invisible to the naked eye, do not leave visible fossil information, therefore making it very difficult to estimate their origins in geological time and predict their characteristics. Here we developed a method that bypasses this challenge by including fossil information for known eukaryotes into the TOL, and applying a novel dating approach that takes advantage of the phylogenetic signatures of eukaryogenesis – specifically that there are branches in both the archaeal and bacterial domains that lead to eukaryotes. Our analyses resolved LUCA at the very earliest periods of planetary evolution on our young Earth around 4.52-4.32 billion years ago. Considering the two primary domains diverging from LUCA, our results time the last bacterial common ancestor (LBCA) to 4.49-4.05 billion years ago and the last archaeal common ancestor (LACA) to 3.95-3.37 billion years ago. Retracing the branches leading to eukaryotes within the Archaea and Bacteria, we estimate the last eukaryotic common ancestor (LECA) to have originated 1.93-1.84 billion years ago. Interestingly, the emergence of LECA in this period highlights how prokaryotes dominated nearly half of planetary history before eukaryotes appeared. Our timeline estimated the ancestor of all Bacteria to be only slightly younger than LUCA. In a separate study included here, we examined the evolutionary histories of all the genes in the genome of a collection of Bacteria to determine the position of the root in the bacterial tree

and predict the gene repertoire of LBCA based on gene presence probabilities. This analysis revealed that LBCA was an already complex organism with a cell membrane, the ability to sense and move in the environment, carbohydrate metabolism, and viral defense mechanisms. Taken together, these data suggest that the earliest periods of planetary evolution following the formation of the Earth were periods of substantial genetic innovation, and hint to an already complex environment in which cellular life was evolving.

The complementary nature of these studies highlights how a reliably resolved TOL can provide a foundation for asking crucial questions about different periods in cellular history and the characteristics of major ancestors at turning points along the evolutionary trajectory. The need for new and advanced tools also ushers in an era of deep investigation into the role of viruses in cellular evolution and how they are compatible within the framework of the TOL. In all, this work provides a greater understanding of past evolutionary processes and new tools that can be applied to address evolutionary inquiries into societally relevant organisms important to human health, global climate change, and food production.

SAMENVATTING

De aarde werd ongeveer 4,5 miljard jaar geleden gevormd en niet lang daarna zal het eerste leven op de planeet zijn verschenen. Als we vandaag de dag onze wereld observeren, zien we dat deze krioelt van een grote verscheidenheid aan organismen, sterk variërend in grootte, complexiteit en levensstijl. Opmerkelijk genoeg zijn alle levende wezens, ondanks deze immense diversiteit in de biosfeer, verenigd door een aantal gemeenschappelijke kenmerken die wijzen op één gedeelde oorsprong van al dat leven. Dat wil zeggen, al het cellulaire leven op aarde zoals wij dat kennen stamt af van één enkele primordiale entiteit die bekend staat als de laatste universele gemeenschappelijke voorouder (LUCA). Wat en wie LUCA was, en hoe het leven op aarde zich heeft gediversifieerd tot moderne organismen, houdt wetenschappers die onze oorsprong willen begrijpen al heel lang bezig.

Een centrale benadering van deze diepgaande vraag heeft zich gericht op het visualiseren van de bekende diversiteit van het leven op aarde binnen een schema dat inzicht geeft in evolutionaire relaties. Rond het jaar 1850 speculeerde Charles Darwin dat alle levende wezens door de tijd heen aan elkaar gerelateerd waren en dat deze verbanden georganiseerd konden worden in een boomachtig netwerk dat meestal een levensboom (Tree of Life, ofwel TOL) wordt genoemd. Hoewel de TOL een bescheiden begin had als schetsen in Darwins notitieboeken, is het sindsdien het middelpunt geworden van de fylogenetica: het vakgebied dat zich bezighoudt met het bestuderen van de evolutionaire geschiedenis tussen organismen. In wezen is de TOL een diagram dat is opgebouwd uit takken en knooppunten die de evolutionaire geschiedenis van het leven weergeven, van onze microbiële voorouders tot het complexe moderne leven, met LUCA aan de basis. Vroege verschijningsvormen van de TOL werden geconstrueerd aan de hand van visuele observaties van morfologische kenmerken, maar de revolutie van de afgelopen eeuw op het gebied van genetic sequencing heeft een enorme diversiteit aan voor het blote oog onzichtbare organismen blootgelegd die een groot deel van de biosfeer uitmaken. De opname van deze lijnen, waarvan er veel nog niet volledig ontwikkeld zijn, in de TOL heeft ons begrip van het evolutionaire verloop van het leven ingrijpend veranderd.

Eeuwen later is de TOL nog steeds een waardevol kader voor het bestuderen van belangrijke aspecten van celevolutie. De eerste moleculaire TOL, dat wil zeggen een boom die wordt afgeleid uit de evolutionaire geschiedenis van gensequenties, werd in 1990 gepresenteerd door Carl Woese en zijn collega's. Zij toonden aan dat al het leven op aarde behoort tot een van de drie grote levensdomeinen, de Archaea, Bacteria en Eukaryota. Belangrijke verschillen tussen deze groepen zijn de celmorfologie en -complexiteit, waarbij Archaea en Bacteria cellen zonder kern hebben en daarom bekend staan als prokaryoten, terwijl Eukaryota of eukaryoten een kern en gespecialiseerde organellen bevatten. Deze aanvankelijke driedeling, of drie-domeinen-boom (3D-boom), bleef vele decennia lang de centrale organisatie van het leven binnen de fylogenetica. Recente updates van deze moleculaire TOL resulteerden in een alternatief evolutiescenario voor de 3D-boom, waarbij in plaats daarvan wordt gesuggereerd dat de cellulaire evolutie het best tot uiting komt in een tweedomeinenboom (2D-boom),

waarbij de Archaea en Bacteria primaire domeinen zijn die afstammen van LUCA. Eukaryoten vormen daarentegen een secundair domein dat is ontstaan door een fusie van voorouderlijke prokaryote cellen via een proces dat eukaryogenese wordt genoemd.

Grotere toegankelijkheid tot genetische informatie en grotere datasets zorgen ervoor dat verbeterde computationele methoden nodig zijn die de gegevens kunnen verwerken en nauwkeurig kunnen modelleren. Over het geheel gezien draagt het werk dat in dit proefschrift wordt gepresenteerd bij aan nieuwe tools en benaderingen die worden gebruikt om de TOL te construeren en leidt het tot kennis die de interpretatie van evolutionaire relaties verbetert in het licht van nieuw ontdekte organismen. De hoofdstukken van dit proefschrift behandelen open en onopgeloste vraagstukken over het evolutionaire verloop van het leven van LUCA tot eukaryogenese, met de nadruk op de opbouw van de TOL en de positie van LUCA daarin, de timing van belangrijke cellulaire overgangen en het metabolisch potentieel van belangrijke voorouders.

Kritische en soms controversiële discussies over de structuur van de TOL duren tot op de dag van vandaag voort, zoals het verschil tussen de 3D- en 2D-boomhypothesen en meer recente discussies over de afstand tussen de twee primaire domeinen, Archaea en Bacteria. Aangezien de prokaryotische lijnen als eerste zich afsplitsten en voortkwamen uit LUCA, is het relevant om hun relatieve posities in de boom goed vast te stellen. Traditionele fylogenetische methoden die een kleine verzameling markergenen (~50) gebruikten om de TOL af te leiden, kwamen uit op een lange tak tussen de domeinen Archaea en Bacteria, wat een aanzienlijke evolutionaire verandering tussen beide suggereert. Een grondige analyse van een verscheidenheid aan markergenen, waaronder een sterk uitgebreide set metabole genen, suggereert echter dat de lengte van deze tak gevoelig is voor de markergenen die worden geselecteerd. Genen met een gecompliceerde evolutionaire geschiedenis, zoals genen die vaak uitgewisseld worden tussen Archaea en Bacteria, trekken de domeinen kunstmatig dichter naar elkaar toe. Mijn analyses laten zien dat zorgvuldige gen- en sequentieselectie nodig is om fylogenetische artefacten te vermijden en dat de best presterende markergenen voor het schatten van de diepe relatie tussen Archaea en Bacteria verticaal evoluerende genen zijn.

Afgezien van de opbouw van de TOL is de timing van belangrijke gebeurtenissen in de celevolutie slecht vastgesteld, met name vanwege de schaarste aan prokaryotische fossielen. Micro-organismen, onzichtbaar voor het blote oog, laten geen zichtbare fossiele informatie achter, waardoor het erg moeilijk is om hun oorsprong in de geologische tijdlijn in te schatten en hun eigenschappen te voorspellen. We hebben hier een methode ontwikkeld die deze uitdaging omzeilt door fossiele informatie over bekende eukaryoten op te nemen in de TOL en een nieuwe dateringsmethode toe te passen die gebruik maakt van de fylogenetische kenmerken van eukaryogenese - in het bijzonder dat er vertakkingen zijn in zowel het archeale als het bacteriële domein die leiden tot eukaryoten. Onze analyses plaatsen LUCA in de allervroegste perioden van planetaire evolutie op onze jonge Aarde, zo'n 4,52-4,32 miljard jaar geleden. Er van uitgaande dat de twee primaire domeinen uit LUCA voortkwamen, zien we

dat de laatste gemeenschappelijke voorouder van bacteriën (LBCA) ongeveer 4,49-4,05 miljard jaar geleden leefde en dat de laatste gemeenschappelijke voorouder van archaea (LACA) ongeveer 3,95-3,37 miljard jaar geleden leefde. Als we de takken die leiden tot eukaryoten binnen de Archaea en Bacteria opnieuw volgen, schatten we dat de laatste eukaryotische gemeenschappelijke voorouder (LECA) 1,93-1,84 miljard jaar geleden is ontstaan. Interessant genoeg laat het ontstaan van LECA in deze periode zien hoe prokaryoten bijna de helft van de planetaire geschiedenis domineerden voordat eukaryoten verschenen. Onze tijdlijn schatte de voorouder van alle Bacteria slechts iets jonger in dan LUCA. In een apart onderzoek, dat hier is opgenomen, onderzochten we de evolutionaire geschiedenis van alle genen in het genoom van een verzameling Bacteria om de positie van de wortel van de bacteriële boom vast te stellen en het genrepertoire van LBCA te bepalen op basis van de waarschijnlijkheid van genaanwezigheid. Uit deze analyse bleek dat LBCA zelf al een complex organisme was met een celmembraan, het vermogen om te voelen en te bewegen in de omgeving, koolhydraatmetabolisme en virale verdedigingsmechanismen. Alles bij elkaar suggereren deze gegevens dat de vroegste perioden van planetaire evolutie na de vorming van de aarde perioden waren van substantiële genetische innovatie en wijzen op een reeds complexe omgeving waarin cellulair leven evolueerde.

Het complementaire karakter van deze studies laat zien hoe een betrouwbaar uitgewerkte TOL een basis kan bieden voor het stellen van cruciale vragen over verschillende perioden in de cellulaire geschiedenis en de kenmerken van belangrijke voorouders op cruciale momenten in het evolutionaire traject. De behoefte aan nieuwe en geavanceerde tools luidt ook een tijdperk in van diepgaand onderzoek naar de rol van virussen in de celevolutie en hoe deze verenigbaar zijn binnen het kader van de TOL. Al met al biedt dit werk een beter begrip van evolutionaire processen in het verleden en nieuwe hulpmiddelen die kunnen worden toegepast bij evolutionair onderzoek naar maatschappelijk relevante organismen die belangrijk zijn voor de menselijke gezondheid, wereldwijde klimaatverandering en voedselproductie.

CHAPTER 1

Introduction

The diversity of life is immense, with organisms colonizing even some of the most extreme habitats on Earth, however it is clear that all life shares common ancestry. Describing, categorizing, and visualizing the relatedness of the biosphere has been a centuries-long endeavor. As one can imagine, our view of life's history has evolved in tandem with advancements in environmental sampling, genome sequencing, and cultivation techniques, among others. Yet, major challenges remain when investigating ancient relationships, including the lack of physical and molecular evidence from the earliest periods of the planet. Universally, all organisms contain some type of genetic information storage molecule (DNA or RNA) that is accessed, encoded, and translated by protein machines to carry out all the necessary functions of the cell. Changes in these genetic components are relics of the distant past and can be used as a window into the earliest periods of biological evolution. In this thesis, I address some of the most enigmatic unanswered questions surrounding the history of cellular life, including the position of deepest split in the evolutionary tree, the timing of cellular evolution, the characteristics of major ancestors, and the metabolic transitions that diversified life as we know it, all through the lens of protein evolution. To fully appreciate this complicated story, I will first outline the origin of life on Earth.

THE ORIGIN OF LIFE

ABIOGENESIS, FROM CHEMICALS TO THE FIRST CELLS

The origin of cellular life is often regarded as one of the most captivating questions in science. Life's origin story can be summarized as a prebiotic to biotic transition involving multiple steps and intermediates ultimately leading to the first forms of cellular life. Cellular life is defined by its shared commonalities, including organized self-sustaining biochemical systems, proteins, and genetic material, all of which are contained within a semipermeable, energized membrane barrier. Self-replication of the genome as well as the cell's structural components are key features of cellular life. The foundational question in life's evolution is how life originated and led to the emergence of the first cells under prebiotic planetary conditions. The transition from precellular chemical entities in the primordial young Earth to bona fide cellular life, including the gradient of evolutionary intermediates, is intimately linked to early geochemistry.

Life's origin story begins around 4.5 Ga during the Hadean eon, a volatile period characterized by rampant volcanism and geochemical cycling inside and on the surface of the Earth. Importantly, abiotic processes in the atmosphere, ocean, and crusts were churning out some of the very basic chemical building blocks required for life, including amino acids, nucleic acids, proteins, metal cofactors, and small hydrocarbons. This was the foundation for the "prebiotic soup" hypothesis, which describes abiogenesis as having occurred in an aqueous environment, such as the primitive oceans or shallow pools, rich in the chemical precursors for life (*1*). Key to this theory was that UV irradiation from the atmosphere would have provided the necessary energy to catalyze the formation of more complex organic molecules that formed life. In 1952, the groundbreaking Miller-Urey experiment provided support for this theory,

showcasing the abiotic synthesis of amino acids and other simple macromolecules from water vapor and chemicals exposed to electrical discharges inside an enclosed system mimicking the conditions of the Archaean environment (*2*). Early hypotheses localized these events in the warm surface of the Archaean oceans or shallow pools atop newly formed continental crusts, but major issues remained related to the concentration of chemical precursors and the ability of lipids to self-assemble into the required compartment in such large and unconfined aqueous solutions. Similarly, UV radiation would have posed a serious challenge to biological molecules and life due to its destructive properties.

The discovery of hydrothermal vents in the late 1970s (*3*, *4*) provided a satisfying solution to these unanswered questions. The first hydrothermal vent systems discovered were found at spreading zones or active volcanoes on the ocean floor, where magma reacts with cool seawater creating a superhot acidic environment, rich in dissolved gasses, small hydrocarbons, sulfides, and metals - a seemingly ideal prebiotic milieu. Despite being shielded from UV radiation and having widespread chemical diversity, super high temperatures at the vents are believed to be biologically limiting and there was no obvious scenario for compartmentalization. The Lost City hydrothermal vent field was discovered in 2000 in close proximity to the Mid-Atlantic Ridge (*5*), shedding light into alkaline vents and providing a plausible alternative scenario for abiogenesis (*6*). Unlike their volcanic counterparts, alkaline vent systems have a higher pH and are not associated with volcanic activity resulting in both lower acidity and temperatures, respectively. The geochemical foundation of alkaline vents is linked to its mineral composition, in the case of Lost City, serpentinization of mafic and ultramafic rock abiotically produces key biological precursors such as hydrogen (H_2), carbon dioxide (CO_2), methane (CH_4), and ammonia, among others (*5*, *7*, *8*). The surface of the calcium carbonate towers at alkaline vents are impregnated with a multitude of tiny interconnected pores through which molecule-rich hydrothermal effluent circulates, providing an ideal nucleation site for encapsulation via the formation of a rudimentary lipid barrier (*9*). The microenvironments that form at the site of these pores may have preceded full cellular compartmentalization by concentrating the necessary chemical precursors together in a confined space. Additionally, abundant H_2 (effluent) and CO_2 (seawater) pools at these vent sites provided adequate energy and chemical conditions for redox-based proto-metabolisms. The narrow mineral channels in the alkaline vents would have enabled the concentration of various chemicals, proteins (*10*), and nucleic acid precursors (Fig. 1) (*11*). Abiotically synthesized lipids (*12*), could have assembled along gaps in the inorganic scaffolds at the pores, concentrating these molecular building blocks into small microcapsules. In time, complete replacement of the scaffold with a lipid bilayer would form a rudimentary membrane preceding the eventual formation of a fully encapsulated vesicle (Fig. 1). Acquiring the energy necessary to “escape” the vent pores and access a renewable energy source was another key transition in the origination of cellular life. The H_2/CO_2 interface in the vent fluids could have provided adequate redox conditions for proto-metabolism akin to the acetyl-CoA cycle (hereafter Wood-Ljungdahl pathway, WLP) (*13*, *14*). The high-energy thioester bond of acetyl-CoA has been proposed to be an early energy currency given such chemical conditions (*15*). Furthermore, producing acetyl-CoA from its H_2/

CO_2 precursors requires no exogenous energy input, and therefore would be a favorable proto-metabolism in the energy-poor primordial pre-cellular system (*16*). To be fully autonomous, the pre-cellular to cellular transition also required access to a self-renewing energy source in order to power biosynthetic pathways. To this end, the primordial membrane barriers in the vents could have exemplified some of the earliest instances of chemiosmosis, where ions flow "down" their chemical gradient across the membrane and release potential energy which can be harnessed to sustain other biochemical processes (*17*). In the pre-cell this may have resembled a proton motive force due to differential pH developed across the primordial membrane. In extant organisms, this proton motive force is instrumental in producing energy in the form of ATP (adenosine triphosphate) via the ATP synthase to support a litany of cellular processes (*18*).

Under such conditions, the pre-cellular to cellular transition could be envisaged as occurring in the small pores, where rudimentary lipid membranes enclosed a concentrated chemical milieu containing all the elements necessary for early biochemical processes. Eventually "escaping" the vent pore, this protocell, although less complex than extant cellular life, represented the earliest *bona fide* cellular organism at the basis of the diversity of cellular life present in our biosphere (*19*) (Fig. 1).

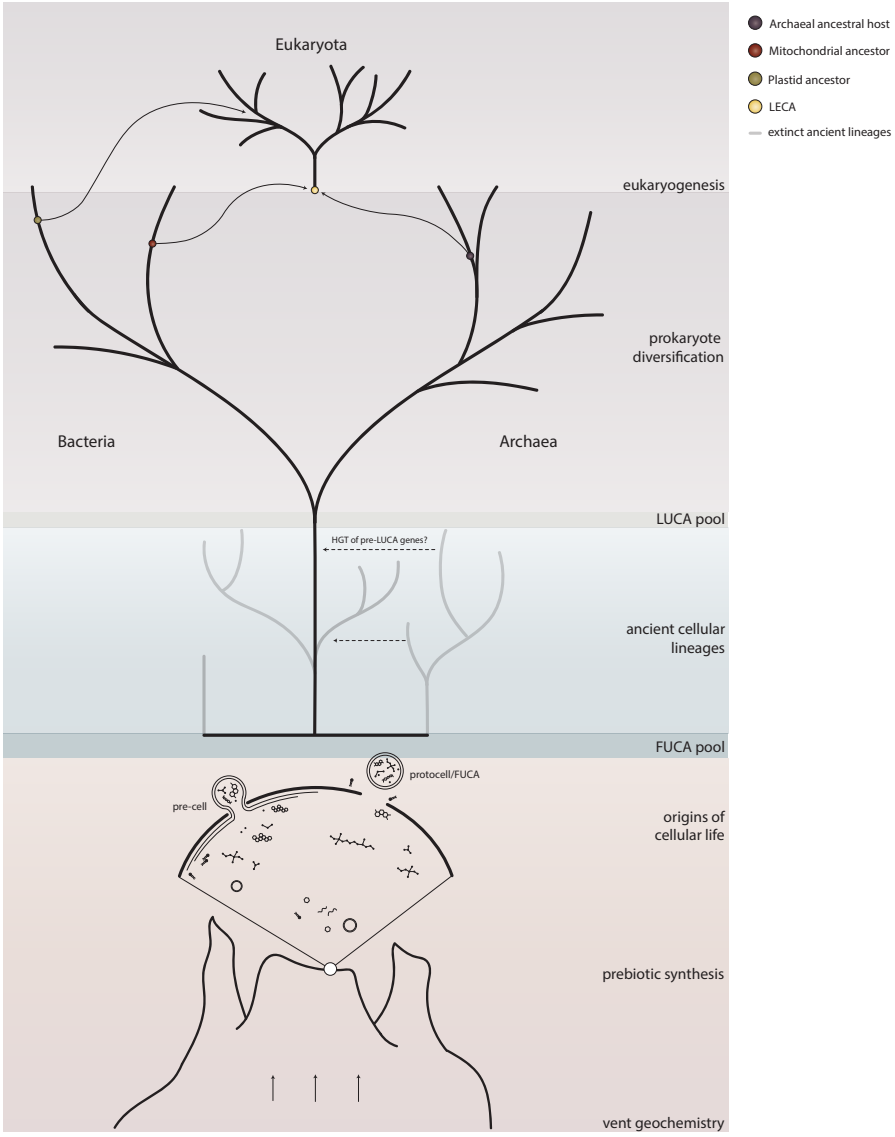


Fig. 1 | Simplified overview of abiogenesis at alkaline vents. Geochemical processes deep within submarine alkaline vent systems are the foundation of this proposed scenario of abiogenesis. In stepwise progression, vent geochemistry gave rise to an active and rich prebiotic chemical environment where small molecular precursors could form from abiotic processes, eventually giving rise to the first protocell or consortium of protocells. The FUCA (first universal common ancestor) pool may have given rise to multiple diversified cellular lineages, which may have gone extinct. A single lineage descending into LUCA (the last universal common ancestor) would be the root of the hypothetical tree of life (TOL), from which the major domains, Archaea and Bacteria diverged and radiated. Eventually, multiple endosymbiotic events between branches of the archaeal and bacterial trees would give rise to the eukaryotes. Ancient extinct cellular lineages are represented in light gray. The archaeal ancestor of eukaryotes is highlighted in purple, the alphaproteobacterial ancestor of the mitochondria is highlighted in red, the cyanobacterial ancestor of the plastid is highlighted in green, and the last universal eukaryotic ancestor (LECA) is highlighted in yellow.

THE LAST UNIVERSAL COMMON ANCESTOR OF ALL EXTANT CELLULAR LIFE ON EARTH

When discussing the stages of life's origins, it is important to highlight a few key concepts. The first cells that emerged at the vents are sometimes loosely referred to as protocells (Fig. 1), a spontaneous self-assembled compartment hypothesized to have had rudimentary RNA (ribonucleic acid) translation processes (*19–22*). Another key stage is the first universal common ancestor (FUCA), which represents the first ancestor of all modern life and any extinct lineages (*22*) (Fig. 1). It is possible that the prebiotic-to-probiotic transition was not a singularity in Earth's history, and cellular life may have been derived but gone extinct on multiple occasions. Presumably, many lineages could have diverged from FUCA without leaving any trace (Fig. 1). However diverse the post-FUCA biosphere may have been, genetic and morphologic commonalities between all extant life show descent from a single ancestral lineage with the entity at its root known as the last universal common ancestor (LUCA) (Fig. 1). LUCA is a hypothetical evolutionary intermediate, representing either a single cellular ancestor or a consortium of ancestors, that bridges primordial cellular life to extant life as we know it.

Different hypotheses have been proposed in the past several decades to describe the nature of LUCA (*23–34*). Carl Woese and George Fox described the hypothetical progenote, which held the same role as LUCA in their interpretation, as cells with pre-prokaryotic (i.e., single-celled microbes) organization (*21*, *23*). More recent work has suggested that LUCA metabolically and morphologically resembled members of the prokaryotic lineages (*32*, *34*), in agreement with earlier proposals (*27–29*, *31*). It would have had a semipermeable membrane serving as both a physical barrier between the genome, proteins, cofactors, and internal biochemistry from the environment, and an energy mediator for chemiosmotic coupling. While ribosomes are confidently traced back to LUCA, the nature of the genomic material has been debated. Earlier evolutionary stages, such as the pre- and protocells, may have propagated from RNA-based genetic systems (*20*, *22*), and the possibility of an RNA-based LUCA has also been suggested (*35–37*). However, the presence of DNA (deoxyribonucleic acid) binding proteins on LUCA's reconstructed genome and the universality of DNA as the primary genetic storage material for extant life, supports a DNA-based genome in LUCA (*32*, *34*).

As is required for all cellular life, LUCA would have been an autonomous self-replicating cell, equipped with a self-sustaining but simple metabolism. It is generally accepted that LUCA was an anaerobe that evolved in a light-poor environment (similar to a hydrothermal vent system), however debate over LUCA's metabolism vacillates between whether the earliest cellular life was autotrophic, heterotrophic, or mixotrophic. Early in the 20th century, Alexander Oparin followed by John Haldane proposed a heterotrophic primordial metabolism based on the consumption of small organic molecules that were formed in the early Earth environment through abiotic processes, the latter of which was the basis for the primordial soup hypothesis (*1*, *38*, *39*). The Oparin-Haldane proposal argued that fermentative metabolisms would have likely preceded autotrophic processes due to their relative simplicity and ubiquity across extant life (*40–42*). Although this proposal is compatible with a low-energy young Earth

ecosystem, where primitive less-complex metabolic pathways existed, it is unclear how these cells would have had access to biochemical precursors. The situation is further complicated when trying to reconcile whether autotrophs would be a required pretext for fermentation based heterotrophs (*43*). Alternative scenarios postulated that the early metabolism of LUCA was autotrophic, based on the conversion of CO_2, which was rich in the early Earth atmosphere and oceans, into organic matter (*13*, *32*, *44–47*). Iron-sulfur catalysts ubiquitous in the early ocean (see above) would have mediated steps in primitive carbon-fixing pathways. Recent analyses trend toward the self-sufficient nature of autotrophic metabolisms, with findings that LUCA was likely an H_2-dependent ancestor involved in the nitrogen cycle, and capable of carbon-fixation via the WLP (*13*, *32*, *34*, *46*). The WLP, an anaerobic carbon-fixation pathway, is frequently implicated as the earliest metabolism due to access to the chemical precursors in the prebiotic environment, and the low energetic threshold (see above). A phylogenetic analysis highlighted the antiquity of the key enzyme of the WLP, the carbon monoxide dehydrogenase acetyl coenzyme A synthase (CODH/ACS), resolving at least four of its primary subunits to LUCA (*48*). However, the authors caution against using the CODH/ACS as a singular marker for LUCA's metabolism as the directionality of the ancestral enzymatic complex is difficult to ascertain from limited data and hold that its presence is compatible with an autotrophic, heterotrophic, or mixotrophic ancestor. Weiss and coworkers, suggest that the resolution of the CODH/ACS in addition to other autotrophic enzymatic complexes, confidently supports the autotrophy of LUCA (*32*, *34*). Additional analyses sampling a larger proportion of the biosphere are necessary to further test these hypotheses.

LIFE'S EVOLUTIONARY HISTORY AND THE TREE OF LIFE

HISTORY

Visualizing life's shared ancestry from LUCA to the current biosphere has long captivated scientists. Systematists built early biological categorization systems primarily from physically observable traits of organisms until the advent of molecular analysis. Swedish botanist Carl Linnaeus formulated *Systema Naturae* in the 18th century (*49*), which classified all plants and animals into groups based on shared characteristics. Going further, he developed a system of ranked hierarchies based on shared morphological and ecological features. Linnaeus' hierarchy provided a foundation for modern taxonomy, a system of classification and categorization of biological organisms. Linnaeus' biological classification system was best represented as a bifurcating tree with the two major branches leading to the plants and animals, respectively. Despite being discovered nearly a century prior (*50*), microorganisms were given very little consideration in this early system as their nature was poorly understood.

In 1859 Charles Darwin published *On the Origin of Species* (*51*), detailing his theory of evolution which is often regarded as the foundation of evolutionary thought. Darwinian evolution describes life's progressive diversification as a consequence of natural selection, where competition and environmental pressure would select for members with certain adaptive

traits (*51*). These changes would not be abrupt or visible in human lifetimes, but instead would result from small cumulative changes over millions or billions of years, a controversial concept at the time. Darwin's groundbreaking "tree of life" visualized biological relatedness in the context of temporal descent, where organisms superseded each other through time based on the principles of Darwinian evolution (*51*). While tree-like diagrams had long been used to demonstrate the interconnectedness of life and the relatedness of individuals (i.e., family trees), Darwin's version attempted to illustrate how this relatedness is modeled through the dimension of time.

The trees depicted in Darwin's journals are regarded as the earliest representations of what is known as a "tree of life" (*51*). At its core, the "tree of life" (hereafter, TOL) is a hypothetical representation of the evolutionary relationships between extant and known extinct organisms and their ancestors. The centerpiece of evolutionary thought, the TOL follows a branching pattern composed of a number of bifurcations from a parent node to descending branches. This branching tree is defined by its "tips" or "leaves" representing extant life, which sit at the end of branches tracing back to a series of bifurcating "nodes" until they reach a centralized "root" of the entire diagram (Fig. 2). The "crown group" is the collection of any ancestor and all of its descendants, while the "stem group" represents the extinct relatives of a given crown group (*52*). The root of the entire tree indicates the shared common ancestor of all organisms depicted in the tree (Fig. 2). In the case of all life, this root would represent LUCA (Fig. 3).

Until this point, the TOL manifested as some combination of plants and animals, as those were the most straightforward to observe and categorize based on related organs or characteristics. Nearly a century later, when microscopic observation became more prevalent, and the characterization of microbial life improved, the *protista* was added as a third kingdom in the initial TOL. A key finding of cytological surveys of microbes revealed a much larger division between organisms. Researchers noted that some cells had membrane bound organelles in the cellular space, most notably being the nuclear membrane that separated the genome from the rest of the intercellular space (*53*). Others lacked this cellular structure and internal compartments. As such, groups with nuclear compartments were classified broadly as *eukaryotes* (true nucleus) while all others were termed *prokaryotes* (before nucleus) (*53*). The prokaryote-eukaryote dichotomy redefined views of the TOL by grouping plants and animals into a much larger group (eukaryotes) which were distinct from single-celled prokaryotes. It wasn't until the late 20th century with the advent of DNA sequencing that more robust groupings appeared and a more accurate view of the TOL emerged (*54*, *55*).

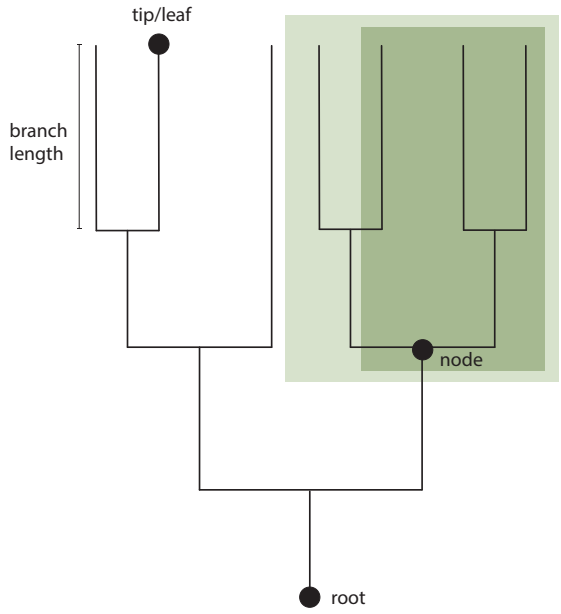


Fig. 2 | The structure of a phylogenetic tree. The main structural components that comprise a phylogenetic tree diagram. Monophyletic groups include an ancestor and all of its descendants (light green). Paraphyletic groups include an ancestor and some, but not all, of the descendants of that ancestor (dark green). The root represents the most recent common ancestor of all entities included in the tree. Nodes are branching points in the phylogenetic tree. Tips/leaves represent individuals, species, populations, or genes.

THE EXPANDING TREE OF LIFE

In the late 1970s, Carl Woese and George E. Fox discovered the Archaea through analyzing the small subunit of the ribosomal RNA (rRNA) (*54*). Despite being indistinguishable under the microscope, they found that prokaryotes were genetically distinct, falling into two major groups, then defined as the *Eubacteria* (hereafter, Bacteria) and *Archaebacteria* (hereafter, Archaea). These molecular analyses also suggested that Archaea and *Eucarya* (hereafter, Eukaryota or eukaryotes) shared a more recent common ancestor with each other than with Bacteria, which was a defining feature of the iconic tripartite TOL published in 1990 (*55*). Based on their molecular analysis of rRNA genes, Woese and coworkers proposed that all life fell into one of the three major domains: Archaea, Bacteria, or Eukaryota (*55*). Technically, the tripartite tree, or three-domains tree (hereafter, 3D tree) (Fig. 3A), is built off the work of Naoyuki Iwabe and coworkers, where pairs of paralogous anciently-duplicated genes, such as the catalytic subunits of the ATP synthase and elongation factors Tu and G, were used to root the TOL (*56*). Their findings indicated a root of the TOL at the deepest split between the branch leading to Bacteria and the branch leading to the sister clades, Archaea and Eukaryota (Fig. 3A). This tree also reflected the evolutionary relevance of their previous findings that Archaea shared similarities with eukaryotes to the exclusion of Bacteria (*54*). In the 3D tree, the Archaea are a sister-group to eukaryotes, suggesting a shared common ancestor at their last ancestral

node. The nature of this important evolutionary split came into question years later when evidence outlined a different scenario to explain the origin of eukaryotes, i.e. the placement of the eukaryotic branch within the Archaea (*57–60*).

At the time, the discovery of the Archaea and resolution of the 3D tree was groundbreaking and it became the center of evolutionary thought and debate for decades to follow (*61*). When looking at the 3D tree, there are some key features to highlight in order to understand the evolutionary relationship between LUCA and its descendants. From LUCA, one branch leads to the Bacteria while the other leads to the Archaea and Eukaryota (Fig. 3A). All crown-group bacteria radiate from a node representing the last bacterial common ancestor (hereafter, LBCA), all crown-group archaea radiate from a node representing the last archaeal common ancestor (LACA), and all crown-group eukaryotes radiate from a node representing the last eukaryotic common ancestor (LECA) (Fig. 3).

Advances in environmental sampling including metagenomic and single-cell genomic approaches (*62–65*) enabled a renewed look at the TOL (*66*), revealing the immense diversity of prokaryotic life resulting from sequencing analyses. This updated tree also resolved two large prokaryotic radiations on either side of the root, the Candidate Phyla Radiation (CPR) (*67*) in Bacteria and the so-called DPANN (named after the first described members: the Diapherotrites, Parv-, Aenigm-, Nano-, and Nanohaloarchaeota) (*68*, *69*) in Archaea. Both the CPR and DPANN are primarily represented by putative symbionts with reduced genome sizes and typically lack or have incomplete biosynthetic pathways considered to be essential (*70*, *71*). Cultivation of members of either group has been difficult, but a few co-cultures have been studied and they typically involve a member of the CPR or DPANN with a more metabolically complete host (*72–77*), indicating symbiotic lifestyles might help supply key biomolecules.

Both the CPR and DPANN were phylogenetically resolved as monophyletic groups basal to their respective domains, which has initially raised questions as to whether early cellular life was less complex (*71*). However, these initial findings are being challenged by more recent phylogenetic analyses that include larger taxon selections and more in-depth phylogenetic analyses (*78*, *79*). Specifically, their proposed symbiotic nature and small genomes have been hypothesized to make them susceptible to phylogenetic artifacts (*80–83*), as is commonly observed in other obligate symbionts and parasites (*84*). Highly reduced genomes often have compositional biases and higher rates of evolution that can lead to long branch attraction (LBA) artifacts, in which long-branching lineages are falsely grouping together (*85*). Additional phylogenetic analyses with sophisticated evolutionary modeling are necessary to adequately position highly reduced symbionts in the TOL.

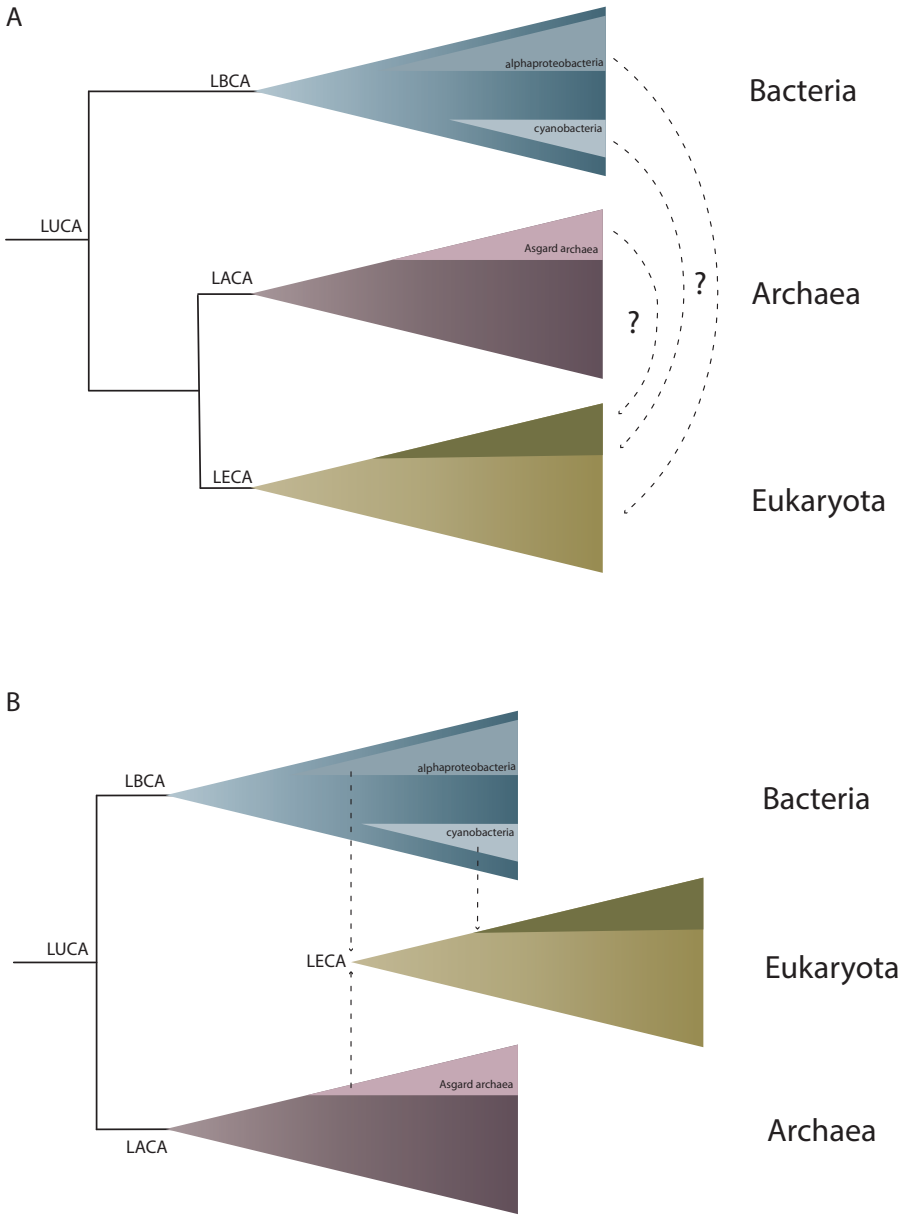


Fig. 3 | **The three-domains and two-domains schematics of the tree of life.** (**A**) A simplified diagram of the three-domains tree (3D tree) proposed by Carl Woese and coworkers (*55*). The 3D tree depicts the three domains, Archaea, Bacteria, and Eukaryota, as monophyletic groups. The deepest bifurcation in this tree is at LUCA, with one branch leading to the last bacterial common ancestor (LBCA) and the other branch leading to the shared ancestor of the Archaea and Eukaryota. Eukaryogenesis scenarios could involve genetic transfer from the prokaryotic domains. (**B**) The two-domain hypothesis (2D tree) proposed in recent years, indicates the same root position at LUCA, however one branch leads to LBCA and the other leads to the last universal archaeal ancestor (LACA), making Bacteria and Archaea primary

domains of life. Later endosymbiotic events involving an ancestral archaeal host with one or two bacterial partners would have given rise to the protoeukaryote and subsequent diversification of the Eukaryota as a secondary domain to the Archaea and Bacteria.

HOW THE ORIGIN OF EUKARYOTES RESHAPES THE TREE OF LIFE

EUKARYOGENESIS

A defining moment in cellular evolution was the origin of the eukaryotic cell, through a transitional phase termed eukaryogenesis. Much of what we know about the visible fossil record suggests that fully fledged eukaryotes did not appear until after the first 2 billion years of Earth's formation. While the consensus is out over the timing of the eukaryotic cell's genesis, it has been estimated to have occurred around 2.0 Ga (*86–88*). The leap from "simpler" prokaryotic cells to the complexity observed in eukaryotic cells confounded the field. How did a cell with such complex structural features evolve? Are they highly derived prokaryotes or something else? Molecular and cytological analyses complicated this story even further, unearthing several conflicting scenarios of eukaryogenesis that are still debated and require further testing (*89*).

A compartmentalized nucleus is typically the central defining characteristic of eukaryotes, differentiating them from prokaryotes. Beyond that, eukaryotes are often larger in size and contain a myriad of cellular structures including membrane bound organelles, vesicles, and a skeletal scaffolding. These regions play an important role in the compartmentalization of cellular functions including DNA replication, protein synthesis, cell division, and metabolism. Interestingly, early cytological studies revealed remarkable similarity between certain eukaryotic organelles and free-living prokaryotes (*90–92*). In the mid-19th century, scientists noticed that the chloroplasts of plant cells bore an uncanny resemblance to free-living cyanobacteria (*90*, *91*). Decades later, the unique morphology of the mitochondria raised the question of its possible prokaryotic origins (*92*). It wasn't until the mid-20th century with advances in microscopy and gene sequencing, that this picture came into full focus. The monumental discovery that chloroplasts and mitochondria had their own native genomes distinct from the eukaryotic nucleus was central to understanding their origins. These new findings shaped Lynn Margulis's controversial theory of endosymbiosis, detailing how the emergence of the eukaryotic cell was defined by, in-part, symbiotic interactions of members from different evolutionary lineages, specifically the prokaryotes (*93*). Under the assumptions of this proposal, eukaryotic organelles, such as the chloroplast and the mitochondria descended from free-living prokaryotes, most likely bacteria, which were engulfed by a proto- or *bona fide* eukaryotic cell. While there was some evidence of horizontal gene transfer in bacteria, the acquisition of an entire cell was a monumental proposal at odds with the idea that life elegantly evolved through vertical descent. Under such a scenario, it would appear that on two occasions (mitochondria and plastid), branches from the Bacteria would have merged with the eukaryotic branch. However, more recent molecular analyses investigating

the placement of the eukaryotic branch in the TOL favor a scenario describing the emergence of the eukaryotic cell from an archaeal host (*94–96*).

THE ANCESTRAL ARCHAEAL HOST AND THE 2D TREE

Increased sampling of archaeal genomic diversity enabled by advances in genome sequencing and the development of more complex phylogenetic approaches culminated in a hypothesis that life's history is better explained by a two-domains tree (hereafter, 2D tree) (*57*, *96–104*). In the 2D tree, the bifurcation from the root node leads to two primary domains, the Archaea and Bacteria (Fig. 3B). Contributions from both primary domains, under the premise that eukaryogenesis involved an archaeal host and one or more bacterial partners, led to the suggestion that eukaryotes are best described as a "secondary" domain (*102*) (Fig. 3B). New evidence supporting the 2D tree was established with the discovery of the Asgard archaea (also known at the Asgardarchaeota (*105*)) through metagenomic approaches (*94*, *95*). Members of the Asgardarchaeota phylogenetically placed as the closest relative to eukaryotes, with eukaryotes emerging as a sister-group to a subclade of the Asgard archaea (*94*, *95*, *106*, *107*). The first discovered Asgard lineage was the Lokiarchaeota, named after the Loki's Castle hydrothermal vent field from which the originating samples were derived (*94*). Several other member clades, also named after other mythic deities, were identified in diverse environments (*95*, *106*, *108*, *109*). Comparative genomic analyses revealed that many members of the Asgardarchaeota contain features previously believed to be unique to eukaryotes and absent in prokaryotes (*94*, *95*, *110*, *111*). Homologs of so-called eukaryotic-signature proteins (ESPs) (*112*) were prevalent across asgardarchaeal genomes (*94*, *95*, *106*, *107*, *113*). ESPs were found to be homologous to eukaryotic proteins involved in key cellular processes including intracellular trafficking, informational processing, and the endosomal sorting complex required for transport (ESCRT) (*114*). In addition, there are parallels between structural components that enable the formation of a putative cytoskeleton (*115*).

Initially, debate raged over the validity of the placement of the eukaryotic branch within the Asgardarchaeota (*116*, *117*), but additional bioinformatics investigations further supported this topology (*96*, *106*, *107*, *118*). While the precise evolutionary position of eukaryotes relative to the Asgardarchaeota is not confirmed, recent studies have confidently placed them within the Heimdallarchaeia, as a sister group of the Hodarchaeales (*106*, *107*). It is important to note that until this point, all speculation of the asgardarchaeal ancestry of eukaryotes was based on bioinformatics analyses, specifically phylogenetic reconstructions using universally conserved marker genes identified on computationally reconstructed genomes. Critiques of such techniques suggested that contamination may have influenced these phylogenetic placements (*119*). However, these doubts have become unlikely with the recent cultivation of the first two cultured members of the Lokiarchaeota, *Candidatus* Prometheumarchaeum syntrophicum (*120*) and *Candidatus* Lokiarchaeum ossiferum (*115*), which provided access to complete genomes from single strains. These studies confidently supported many of the physiological and metabolic proposals made using reconstructed genomes from bioinformatics analyses and confirmed the genuine presence of ESPs. Microscopic observations reveal that these

asgardarchaeal representatives contain cellular structures, including protrusions (*120*) or an actin cytoskeleton (*115*). If present in the archaeal ancestor of eukaryotes, these could have been instrumental in mediating physically intimate relationships between the archaeal host and bacterial partner giving rise to eukaryotes through symbiotic interactions. In time, more closed genomes of members of the Asgard archaea, including members from the *Heimdallarchaeum* (*113*) and *Odinarchaeum* (*121*), were obtained, and in combination with cellular visualization and biochemical assessment, have confirmed that extant Asgard archaea appear to have a number of complex cellular features in common with eukaryotes. A recent analysis found that gene duplications greatly expanded the proto-eukaryotic genome, with genes from the asgardarchaeal ancestor undergoing the most duplications (*122*). Together, this suggests that the last archaeal ancestor of eukaryotes may have already possessed a degree of cellular complexity, key to the evolution of eukaryotic cells.

MITOCHONDRIAL AND PLASTID INHERITANCE

In light of this recent work, we must also consider the bacterial partner(s) that shaped the origins of eukaryotes. Endosymbioses leading to the mitochondrial and plastid organelles are central elements of eukaryogenesis. In 1985, Yang and coworkers used 16S rRNA gene sequencing to show bacterial origins of the mitochondria, proposing that mitochondria were derived from the alpha subdivision of the then-called "purple-sulfur" bacteria (*123*). Genomic and phylogenetic evidence has confidently placed the mitochondrial ancestor amongst the Alphaproteobacteria (*124*) but the specific phylogenetic position of this relationship is debated due to the tendency of mitochondria to cluster with the Rickettsiales, common intracellular parasites (*125–129*). Recent phylogenetic analyses have indicated that the affiliation of the mitochondrial branch with Rickettsiales most likely represents a phylogenetic artifact and instead placed mitochondria as sister to most Alphaproteobacteria (*130–132*). Plastid endosymbiosis is another complicated event as there can be serial inheritance of photosynthetic bacteria from different stems of the cyanobacterial tree (i.e., the primary plastid) (*133*) and lateral transfer of already-established eukaryotic plastids (i.e., secondary and tertiary plastids) that define the evolution of this organelle (*134*). An expanded genomic and phylogenetic analysis of genes in the plastid and native genomes of plastid-bearing eukaryotes revealed an early cyanobacterial origin for the ancestor of plastids, *Gleoemargarita lithophora* (*135*).

Since both organelles play crucial metabolic roles in their respective cells (i.e., plastids are only found in cells of phototrophic organisms), the relationship of host and symbiont was likely linked to biochemical constraints and conditions of metabolic exchange. Several recent review articles outline various hypotheses that have been put forward to explain eukaryogenesis in light of recent data (*89*, *136–139*). While these hypotheses differ in their predictions regarding the nature of interactions, the number and identity of partners involved, the core biochemical conditions driving such interactions, and the relative timing of proto-mitochondrial acquisition, these endosymbiotic proposals typically involve an ancestral archaeal host with one or more bacterial partners. Many models of mitochondrial evolution

are syntrophic or symbiotic models where the alphaproteobacterial partner in some way stabilizes redox conditions in the host cell. The hydrogen hypothesis describes the interaction of a H_2-producing alphaproteobacterium with a H_2-dependent archaeon (*140*). Whereas the syntrophy hypothesis posits that a sulfate-reducing deltaproteobacterium engulfed a H_2-producing archaeon followed by the uptake of an alphaproteobacterium, which later became the nucleus and mitochondria, respectively (*138*, *141*). The reverse-flow model suggests that the alphaproteobacterial partner served as an electron sink for the archaeal host (*120*, *142*). Another scenario is illustrated by the phagocytosing archaeal model where a free-living mitochondrion was suggested to have been engulfed by an archaeon and eventually retained function inside that cell (*143*). Open and unresolved questions related to the timing of proto-mitochondrial acquisition (i.e., early, intermediate, and late) relative to other features of cellular complexity evolved independently by the host (*89*, *128*, *144*), remain. While early phagocytosis models favored an elaborate proto-eukaryote as host, therefore representing mitochondria-late scenarios, the original hydrogen hypothesis is generally seen as a mitochondria-early scenario (*89*). The discovery of the Asgardarchaeota and more recent phylogenetic evidence, are more in line with so-called mitochondria-intermediate scenarios, in which the archaeal host had already evolved a degree of cellular complexity (*94*, *122*, *144*, *145*).

Taken together, there is strong genomic evidence of endosymbiotic events leading to both the mitochondria and plastid, but more research is needed to conclusively determine the ancestor of either group and the possible mechanism of acquisition. Eukaryogenesis is a defining moment in cellular evolution and the history of life on Earth. The myriad of biotic precursors involved in this process emphasizes a key evolutionary feature of eukaryotes – they contain physical but more importantly, genetic information from either side of the root of the TOL (*89*). This is crucial when addressing questions regarding the deep evolutionary history of not just eukaryotes, but also its prokaryotic counterparts, and requires thoughtful application of genetic tools and methods that account for the nature of eukaryotes.

HOW DO WE STUDY THE EVOLUTION OF CELLULAR LIFE?

While Darwin philosophized over the idea that all organisms were related and descended from a common primordial ancestor, he had little empirical evidence to support such claims. Around the same time, Gregor Mendel's research on pea plants defined the concept of heritability, that traits were inherited from parent to offspring by way of genes (then termed factors) (*146*, *147*). The term "gene" has two definitions: the Mendelian definition centers them as the primary unit of heredity, whereas the molecular definition describes them as segments of genetic material that encode for a protein product necessary for cell function (*148*). For all organisms, DNA is the genetic material encoding all genes, where RNA serves as an intermediate informational messenger preceding protein synthesis. Proteins, the molecular machines that drive biochemical reactions, are most commonly produced from the information stored in genetic material. The relationship between the genetic information

and the physical manifestation of traits in the organism is best explained by the relationship between an organism's *genotype* and *phenotype*. In essence, the underlying genetic code (genotype) determines the physical characteristics (phenotype) of an organism. Insight into genes, how they evolve, their inheritance patterns, their distributions, and overall evolutionary histories can shed light into evolution of life and metabolic innovation.

GENES

The passage of genes from parent to offspring is a fundamental principle in biology, however it is not as straightforward of a process as it may appear. When DNA is copied and integrated into daughter cells, random changes in the sequence, or mutations, may be introduced. The degree to which mutations impact the protein sequence and downstream gene product generally depends on the location of the mutation, and can be beneficial, deleterious, or neutral (*149*). These sites of genetic variation are molecular measures used to estimate evolutionary change through time (*150*, *151*). Traditional notions of gene mutation under Darwinian principles have shifted to accommodate the modern complexity of genomics-based inquiry and has expanded to include the following principles of gene variation (*152*, *153*). Substitutions in nucleotide bases is one of the more basic elements of mutation, where any nucleotide may be swapped for another with an added complication being that the likelihood of certain swaps are higher than others. Gene duplication results in two gene products from a duplication of an entire gene, but the downstream changes to the duplicated gene(s) are crucial to understanding how this mode of genetic change influences evolution. In some cases, one gene product gains a new function and diverges from its duplicated counterpart. Likewise, the duplicated gene can have a loss of function, and either be lost from the population due to drift or coevolve with its counterpart due to subspecialization (*152*). The latter scenario is the central principle of evolution of the catalytic subunits of the ATP synthase headpiece complex (*154*, *155*). In other cases, both duplicated products may change function resulting in novel subspecialization of the new genes. Importantly, gene duplications believed to have occurred in or before the LUCA have immense impacts on our understanding of evolutionary history, such as the catalytic and non-catalytic subunits of the ATP synthase heterohexameric headpiece or elongation factors Tu and G, which have been used to root the TOL (*55*, *56*, *154*). The loss of genes can also be evolutionarily innovative, as the loss of one (or more) genes can lead to speciations. Similarly, novel genes can arise from previously noncoding regions of DNA. Finally, horizontal gene transfer (HGT) describes the lateral sharing of genes between different organisms. Originally identified in bacteria (*156*), recent evidence has shown this is a common mechanism of gene flow across the TOL, including with and among eukaryotes (*139*, *153*, *157–164*).

PHYLOGENETICS

Signatures of gene mutation, duplication, transfer, origination, and loss on the genomes of extant organisms can be used as a window into early evolution across the TOL (*152*). Visualizing these variations is the foundation of the field of modern phylogenetics. Phylogenetics is the branch of biology concerning the evolutionary relationships between organisms. Phylogenetic study in the 19th century was built on comparisons of morphological and ecological traits of

organisms, based on principles proposed by Linnaeus, Darwin, and others (see above). Modern phylogenetics is based on molecular sequence data, believed to be more robust and reliable than morphology-based interpretation. Nonetheless, these relationships can be visualized relative to each taxon or gene in a phylogenetic tree, which is a branching diagram composed of tips/leaves (extant taxa), branches, nodes, and a root (see above) (Fig. 2). Crown groups, an ancestor and all their descendants, are considered monophyletic. In contrast a paraphyletic group includes an ancestor and some, but not all, of its descendants - in some cases subgroups are found at different positions in the tree.

The process of selecting molecular markers for phylogenetic analyses depends on whether one is interested in the history of a specific gene family or aims to resolve the evolutionary relationship of organisms relative to each other. In the latter case, generally more information than encoded within a single gene is required. To be able to determine phylogenetic relationships, the analyzed genes have to be homologs that derive from a shared ancestral sequence (*152*). There are two main examples of homologous genes; orthologs originate from a single ancestral gene diversified by speciation, whereas paralogs are genes resulting from duplications within a taxon (*152*, *165*). Gene families that can be used to infer a TOL need to be single-copy, vertically transferred, and universally shared across all domains of life. These criteria generally leave only a small number of gene families (fewer than 50) suitable for inference of the TOL, which only represent a tiny fraction of an organism's genome (*166*). Studies have attempted to use expanded marker gene sets (*167*) to overcome such limitations and infer the TOL using a more comprehensive set of genetic information. However, gene discordance (i.e., gene duplication, transfer, and loss, etc.) can complicate the phylogenetic inference (*168*). Supertree methods can account for incongruencies in gene tree history by inferring a single species topology from a combination of pre-computed single gene phylogenies (*169*, *170*). This method is high-throughput and can be applied to very large datasets.

Detecting homologs of conserved marker genes and distinguishing between true orthologs and paralogs in an organism's genome is often approached by comparing the sequence identity of these homologs to gene sequences in curated databases. The basic local alignment sequence tool (BLAST) (*171*) is instrumental in this process. A more sensitive method of detecting homologs relies on statistical modeling that better reflects the evolutionary processes underlying the evolution of the marker gene. Hidden Markov models (HHMs) can be manually curated from a small collection of homologs or accessed from verified databases that have already generated sets of HMMs, which can be used to detect homologs based on compatibility with the HMM profile (*172*). HMMER is most often used to execute such homology searches (*173*, *174*).

Once sequences have been selected, they are aligned into what is called a multiple sequence alignment (MSA). The MSA is a matrix built off the homology of nucleotides or amino acids of different sequences based on common patterns. Different methods exist to obtain alignments

that maximize certain attributes of the sequences that reflect evolutionary relationships. Different algorithms of homology detection can be used to increase confidence in the prediction of the alignment. A necessary step in this process is visualizing the alignment for quality control. Poor quality regions and gaps are often trimmed out of the MSA to ensure that poorly aligned segments do not bias the downstream phylogenetic inference. One can manually inspect homologous regions and gaps in order to identify sequences that might need to be removed or to remove sites with questionable histories in order to create a more reliable alignment. To ensure reproducibility, especially when larger sets of sequences are analyzed, it is important to rely on software to assist in the process of trimming questionable alignment regions based on a set number of parameters. However necessary, the trimming process must be carefully applied as the removal of too many sites can compromise the phylogenetic signal. In some cases, more relaxed trimming methods that conserve more positions may be more suitable (*175*). The limited resolution of single-gene trees poses several challenges when trying to reliably infer and interpret deep evolutionary relationships between different taxa. Using multiple genes to infer a phylogeny provides additional information from which to robustly inform relationships, a process called concatenation. Concatenated protein alignments are produced from concatenating multiple single-gene alignments into a supermatrix, which results in many more amino acid sites to examine evolutionary change. This concatenated alignment is larger but treated the same as single-gene alignments in the downstream processes.

A statistical framework for evolution is then applied to this sequence alignment to infer a phylogenetic tree. Modern methods of phylogenetic inference rely on maximum likelihood (ML) (*176*) or Bayesian (*177*) statistical estimation (*178*, *179*). The concept of *likelihood* describes the probability of observing data given a phylogenetic tree and model of evolution (*178*). ML estimation is an extension of this principle, defined as a statistical method that optimizes parameters given by a tree and model that best explains the observable data (*178*, *180*), ultimately producing a phylogeny with the highest likelihood based on the data provided. Bayesian methods differ from ML in that probabilities for the same likelihood conditions are estimated based on prior knowledge of the distribution of the observable data (*178*, *179*). In essence, the probability of the likelihood using Bayesian statistical inference is weighted by the conditionality of other prior events or information.

A model of sequence evolution describes the underlying evolutionary processes acting on the sequence data. These models are typically represented as a two-dimensional matrix containing probability scores of the exchange of one character (nucleotides or amino acids, 4 or 20 matrix categories, respectively) to another. They are applied under both ML and Bayesian probabilistic methods and can vary greatly in complexity (*178*). DNA substitution models take into account the substitution rate of the exchange of one base to another and the base frequency. In the simplest example, the Jukes-Cantor (JC) model of evolution considers that all possible base substitutions have equal probability and that the base frequencies are equal (*181*). More complex nucleotide models, such as the General Time Reversible (GTR) model,

considers unique substitution rates and unequal base frequencies (*182*). Similar models, such as Le and Guascual (LG) apply to protein datasets (*183*).

Additional criteria must be considered for protein evolution, where regions within a protein sequence may evolve faster or slower depending on protein structure and function. Considering the degenerate nature of the genetic code, the third codon position in protein-coding sequences can mutate at a different rate than the first or second positions (*178*). Furthermore, highly conserved catalytic regions might have different evolutionary rates to non-catalytic or structural parts of the protein depending on evolutionary constraints. This concept is known as rate-heterogeneity and can be accounted for using more sophisticated modeling such as gamma distributions or mixture models in order to make more realistic predictions of biological change over time. A discrete gamma distribution model contains categories of rates, all of which have equal probability, which are used to approximate the likelihood of specific sites compared to the mean of each category (*184*). On the other hand, protein mixture models integrate additional amino acid rate substitution matrices to better explain protein evolution, compared to a standard single matrix. Common models use 10-60 different rate matrices (termed C10-C60) to model sites along a protein sequence (*185*). Model selection can depend on how suitable the data is to particular statistical frameworks within the model and mode of inference. A number of softwares can select the best model fit to the provided data selected based on statistical criteria that quantifies error and relative quality of the estimation (e.g., see Akaike information criterion and Bayesian information criterion) (*186*).

Several different tree topologies are then computed and cumulatively assessed along with other parameters (i.e., branch lengths and model) to identify the scenario that optimizes the defined conditions, ultimately resulting in the tree with the highest likelihood (see ML estimation above). Phylogenies are estimations of evolutionary relationships based on statistical frameworks, therefore confidence in the tree can be statistically tested. Bootstrapping is a computational technique based on resampling and replication of phylogenetic inferences and was initially used to determine confidence in observed phylogenetic relationships (*187*). The bootstrap estimate is a value out of 100 representing the number of times the same branch is observed when the data is repeatedly, randomly resampled and a phylogeny is regenerated from any subsample. Robust phylogenetic analyses typically employ a bootstrap approximation of 500-1000 replicates, and phylogenetic programs contain a variety of methods that enable computing these values while reducing computational burden (*188*). Tree confidence can also be measured according to likelihoods and probabilities using a variety of statistical tests. Felsenstein's bootstrap approximation (*187*) was the foundation for the bootstrap selection probability (BP) topology estimate. While this work was seminal in the field of phylogenetics, several studies have shown BP may not be the most suitable tool for modern datasets due to biases and poor fit to larger samples (*189*, *190*). As a result more robust statistical tests were developed to assess tree topology (*178*, *191–193*). Currently the most reliable method is a more complex statistical test termed the Approximately Unbiased (AU) test, which attempts to reduce bias in a multiscale bootstrap technique (*193*).

GENE-TREE SPECIES-TREE RECONCILIATIONS

Robust phylogenies are important for any downstream analyses and interpretations that rely on an accurate understanding of evolutionary relationships, such as gene-tree species-tree reconciliation analyses. Gene tree-species tree reconciliation analysis, model (reconcile) the most likely evolution of gene families, under a given species tree scenario. Specifically, Amalgamated likelihood estimation (ALE) is one program used to model these reconciliations and, in doing so, estimates the probability of gene duplication, transfer, origination, and loss events across the tree (*194*). This reconciliation process has proven useful in estimating the vertical and horizontal contributions to evolution, applying outgroup-free rooting methods to avoid phylogenetic artifacts (*195*), and reconstructing the enzymatic repertoire of key ancestors in a species tree (*107*, *195*). The probabilities of gene family presence at different positions in the species tree is useful to understand the metabolic capabilities of ancestors and lineages and provides insight into major evolutionary transitions. In addition, reconciliations provide a statistical framework for testing root hypotheses in the absence of an outgroup - here any hypothesized root position can be tested and assigned a probability.

MOLECULAR DATING AND CROSS-BRACING

In addition, a reliable phylogeny is crucial to molecular clock analyses, as the appropriate application of fossil ages and age ranges can affect any evolutionary outcomes if taxa (or lineages) are poorly placed in a species tree. Providing a timeline for major cellular transitions and speciations, enables deeper understanding of life's history and ecological impact. The use of molecular sequence data to estimate the timing of major events in cellular history, or molecular dating, is a valuable tool in evolutionary biology (*196*, *197*). The premise of molecular dating is based on the proposal that time between speciations can be measured by assessing differences in molecular sequences (*197*). Molecular dating requires a reliable phylogeny, a model of evolution, a model of rate variation, and a calibration of node ages (*196*, *198*). Node calibrations typically correspond to a particular date, age, or age range (minimum and maximum age constraints), which are often derived from the fossil record or geological data (*196*). Molecular clocks can be strict or relaxed, in the former a single constant is used whereas in the latter the rate of variation differs along different branches. More sophisticated methods of molecular clock analyses, especially those that model rate variation can be computationally demanding, making analyses costly in time and resources.

Dating deep events in the TOL can be challenging because of the paucity of fossils in prokaryotes. To date, there are no archaeal fossils and only a few fossils in Bacteria, mostly within the Cyanobacteria. Similarly, despite advances in more recent molecular clock models, they do not confidently model rate variation on deep branches. Therefore, very sophisticated methods are required to time events early in prokaryotic history. One way to overcome this limitation is to apply a relative constraint, where relative node order is incorporated into clock analyses to constrain ages (*198*, *199*). An additional measure would be to identify marker genes where multiple nodes correspond to the same evolutionary event in their gene tree, a trait of duplicated genes and genes inherited by eukaryotes from their prokaryotic ancestors

(i.e., horizontal transfer of genes). Equivalent speciation nodes on a given phylogeny can be "cross-braced" (*200*, *201*) together, and age calibrations can link those braced nodes in time. In combination with a relative constraint, such as the mitochondria necessarily preceding the plastid in evolutionary time, together provide an added temporal dimension to molecular dating to overcome previously mentioned limitations.

RIBOSOMES AND THE ATP SYNTHASE AS A WINDOW INTO LIFE'S EVOLUTION

Ribosomes and the ATP synthase have unique evolutionary features that make them valuable molecular markers for analyzing key events in cellular history. The ribosome is one of the most consequential enzymatic complexes to have evolved. Found in all cells, ribosomes synthesize proteins by polymerizing amino acids together based on a set of instructions encoded by RNA. Protein synthesis is a fundamental process for cellular organisms and is universally conserved across all extant taxa. As a result, the evolutionary mechanisms underlying the history of the ribosome is a unique window into what the early genetic and enzymatic landscape may have looked like on primordial Earth. RNA transcripts have unique features, including their ability to take on secondary structural conformations and act both as genetic material and catalysts, which are believed to have been important in the primordial environment and for early cellular life. The autocatalytic properties of RNA transcripts were confirmed in 1982 (*202*) through a series of self-splicing experiments. The ability of RNA molecules to act as both genetic material and biological catalysts is the centerpiece of the RNA world theory, the idea that RNA-based genetic systems dominated the prebiotic replicator space and preceded DNA (*19*). Ribozymes, as they became known, are RNA molecules with enzyme-like catalytic properties, such as self-splicing, gene, expression, and amino acid polymerization.

Two major complexes comprise the ribosome, a small- and large-subunit, which together form the translation machinery. Ribosomes are made up of anywhere from 55-79 individual ribosomal proteins, with eukaryotes having more due to their larger complex (*203*). Ribosomal proteins are valuable molecular markers because the ribosome is universally conserved across all life. In addition, since eukaryotes retained certain cellular features of their prokaryotic ancestors via endosymbiosis of the mitochondria and plastid, they can have up to three distinct pools of ribosomes conserved between the mitochondria, plastid, and nucleus. Therefore, one can retrace the evolutionary histories of each eukaryotic pool of ribosomes in order to explore deeper events in the prokaryotic domains.

Like the ribosome, the ATP synthase is another crucial enzyme implicated in life's origins. The ATP synthase is a rotor-stator complex central to energy generation via the synthesis of ATP from ADP and inorganic phosphate (*204*, *205*). The presence of an ATP synthase in the pre-cells, protocells, and/or LUCA at the very least, is not unreasonable given its function and distribution across extant life. Similarly, many theories of abiogenesis speculate the ATP synthase may have already been present in the primitive membrane to relieve ion gradients naturally building across the barrier (*32*, *46*, *47*, *205*, *206*). Unsurprisingly, the ATP synthase is universally conserved across all domains of life, which may reflect how crucial ATP synthesis

was for life's evolution. Briefly, in extant organisms there are three main types of ATP synthases, the F-type found primarily in Bacteria, the A-type found in Archaea, and the V-type found in Eukaryotes. The A- and V-type ATP synthases belong to a larger family termed the A/V-type ATP synthases. The F-type appears to be the structurally simplest version of the enzyme, lacking the peripheral groups typical of A/V-types (*204*, *207*).

Regardless of type, these macromolecular machines have a conserved structure including a hydrophilic catalytic headpiece, the site of ATP synthesis, which is connected to a transmembrane ion-translocating ring (*204*, *207*). Functionally, the ATP synthase couples the translocation of ions (protons or sodium) across the membrane to ATP synthesis, making it a valuable in the pre-cellular to protocell transition, as it could take advantage of the naturally formed proton motive force to generate energy for other biochemical processes. The sodium-proton dichotomy is believed to also influence evolution of the ATP synthase in that early membranes may have been leaky to either ion leading to dependence on the other (*31*, *206*, *208–210*). This has not been confirmed experimentally and no true pattern for proton versus sodium dependence has been verified, therefore further research is necessary to understand this early invention. In the porous alkaline vent, this redox activity would have relieved the thermodynamic constraints of a naturally accumulating proton motive force. Therefore, despite its structural complexity, the ATP synthase may have been an elegant solution to a lingering problem in balancing thermodynamic constraints of the chemical system. Beyond resolving the proton gradient, the ATP synthase contributed to another energetic breakthrough by simply producing ATP, the primary energy currency of life. With a functioning ATP synthase, the protocell could have taken the energetic leap it required to fully escape the vent environment and become a self-sustaining entity, as deriving energy could now be done by tapping into an ion-gradient-energized membrane.

Similar to the ribosome, the ATP synthase is a useful molecular marker for examining evolutionary history from the time of LUCA to now. The primary catalytic sites in the headpiece that mediate ATP synthesis from its precursors are the consequence of a very ancient gene duplication followed by a loss of function in one subunit (*56*, *154*, *155*). Evidence of this gene duplication is apparent in phylogenetic comparisons of gene sequences collected from Archaea, Bacteria, and eukaryotes (*155*, *200*). Each paralog (from the gene duplication) can be used to outgroup-root the other, which has historically been applied to root the TOL between the Bacteria and the clade containing Archaea and eukaryotes (*55*, *56*, *154*). Eukaryotes inherited ATP synthase complexes from both their prokaryotic ancestors, with the A/V-types in vacuoles and the F-type on the mitochondria and/or the plastids. Therefore, like with the ribosome, eukaryotes have 2-3 (depending on the presence of a plastid) distinct ATP synthase pools, which can also be used to retrace early evolutionary transitions.

In general, both the ribosomal proteins and ATP synthases have long been included in universal marker gene sets for inferring the TOL, due to their conservation across the three domains. The iconic TOL inferred by Hug and coworkers was inferred using 16 ribosomal proteins (*66*).

The endosymbiotic inheritance of the ribosome and ATP synthase, as well as the ancient duplication in the latter, results in the same speciation event(s) being represented multiple times in their respective gene trees. This node equivalency can provide a constraint for age (fossil) calibrations that are typically used in molecular clock studies to date evolutionary timelines. This principle has previously been applied with the ATP synthase to time the time the mitochondrial and plastid endosymbiosis events to the late Proterozoic (*200*). Taken together, the ATP synthase and ribosomal proteins can provide valuable insight into the earliest periods of cellular and metabolic evolution from LUCA to extant life.

THESIS SCOPE

Retracing the major events early in life's evolution is challenging and requires careful and methodical inquiry. This thesis aims to address unresolved key questions regarding deep cellular evolution from the earliest node in the tree of life, representing LUCA, to the divergence and radiation of Archaea and Bacteria, endosymbioses contributing to eukaryogenesis, and the structure of the TOL. To this end, I applied diverse bioinformatics approaches including phylogenetics (single genes, sets of genes, all genes on a genome), gene tree-species tree reconciliations, molecular dating, metabolic reconstructions and ancestral reconstructions. Through the lens of enzyme evolution, I examine the shape of the tree of the primary domains of life and their evolutionary distance (Chapter 2), the timing of cellular evolution and the history of the ATP synthase (Chapter 3), and the nature of the LBCA (Chapter 4). Chapter 5 is a literature review on current perspectives on the TOL including the evolutionary impact of the enigmatic Virosphere. The final chapter (Chapter 6) discusses the diversity and global impact of the Archaea.

CHAPTER 2: THE DEEPEST SPLIT IN THE TREE OF LIFE

In **Chapter 2** I examine the evolutionary proximity of the two primary domains. Universal core marker gene sets shared across Archaea and Bacteria typically contain up to ~50 genes, usually associated with informational processing machinery (e.g., transcription, translation, genetic processing), but representing a small fraction of genetic information from an organism's genome. To overcome this limitation, a collection of 381 marker genes, including many metabolic genes, were used to infer a phylogeny of 10,575 Archaea and Bacteria by Zhu and coworkers (*167*). While the phylogeny was inferred using a supertree method in ASTRAL (*211*), traditional marker gene concatenation was applied to infer branch lengths. As a result, their analysis resolved a short evolutionary distance between Archaea and Bacteria, which conflicts with results from traditional analyses. To investigate this discrepancy I compared different marker gene sets of varying size and gene type (i.e. ribosomal, non-ribosomal, and core marker genes) to this larger expanded set (*167*) of genes. I used two different methods to assess the reliability of marker genes in the expanded set and our curated test set: a manual inspection of phylogenies and a quantification of taxonomic splits. First, I manually inspected gene phylogenies to see if they recovered the reciprocal monophyly of the domains, as well as

for the presence of HGT and paralogous gene sequences. Inferring a concatenated phylogeny using genes with complicated histories and non-vertical inheritance, can cause phylogenetic artifacts. In the second method, I ranked the best and worst performing marker genes based on how well they recovered established sisterhood relationships. Results demonstrated that phylogenetic analyses are sensitive to marker gene selection. The inclusion of paralogs rather than orthologs and intradomain HGT results in a phylogeny with an artificially shortened branch between Archaea and Bacteria. Statistically, the best performing markers resolved a tree with a long interdomain branch between the primary domains and recovered the positioning of key groups consistent with recent analyses.

CHAPTER 3: TIMING CELLULAR EVOLUTION AND THE HISTORY OF THE ATP SYNTHASE

In **Chapter 3** the timing and evolutionary history of the ATP synthase was investigated based on a dated cross-braced TOL. Dating deep evolutionary events is complicated by the paucity of prokaryotic fossils and the failure of molecular clocks to properly model rate variation. In turn, adding eukaryotes, for which fossils are available, to phylogenetic analyses is key. For this, it is crucial to use phylogenetic marker genes that are not only shared between Archaea and Bacteria but are also present in eukaryotes and inherited from their archaeal and bacterial ancestors, respectively. Importantly, eukaryotes have inherited up to three sets of ribosomes and ATP synthases from their prokaryotic ancestors. The ATP synthase core function is enabled by two headpiece subunits, termed catalytic and non-catalytic, that mediate the synthesis of ATP from ADP and inorganic phosphate. These two subunits are the product of an ancient duplication, believed to precede LUCA. As a result, speciation events across the Archaea, Bacteria, and eukaryotes are represented twice in a phylogeny of these ATP synthase subunits. Taken together, gene duplication (ATP synthase) and endosymbiosis (ATP synthase and ribosomal proteins) results in multiple nodes of the respective gene trees corresponding to the same evolutionary event. This node equivalence was the foundation for the cross-bracing method which we combined with a relative time constraint, to propagate eukaryotic fossil ages into poorly constrained regions deep in a TOL inferred from ribosomal proteins and a gene tree of the key subunits of the ATP synthase. Dating analyses revealed that LUCA lived around 4.52-4.32 Ga and LBCA lived from 4.49-4.05 Ga, both preceding LACA which was estimated to live around 3.95-3.37 Ga. The timeline for cellular evolution resolved from this dated TOL provided an absolute timing for the ATP synthase, revealing that the ancestral duplication of the catalytic and non-catalytic subunits (4.52-4.46 Ga), as well as the speciation of the two major types of ATP synthase, the F- and the A/V-type, overlapped or preceded the timing of LUCA. In combination with phylogenomic and phylogenetic analyses of the catalytic and non-catalytic subunits of the ATP synthase, multiple evolutionary scenarios are presented where LUCA either 1) contained a rudimentary F-type ATP synthase, with gradual evolution of the A/V-type along the archaeal stem and a transfer into LBCA or 2) that LUCA already contained both the F- and A/V-type ATP synthases, the former of which was lost along the stem to LACA with a very late transfer of F-type ATP synthases into the Methanosarcina.

CHAPTER 4: THE BACTERIAL PHYLOGENY AND THE LAST BACTERIAL COMMON ANCESTOR

In **Chapter 4** advances in gene tree-species tree reconciliations were instrumental in outgroup-free rooting of the bacterial tree, quantifying the vertical and horizontal contributions to bacterial evolution, and reconstructing the putative enzymatic repertoire of LBCA. The bacterial phylogeny convincingly resolved a deep split between the Terrabacteria and Gracilicutes, with the CPR being a derived lineage within the Terrabacteria, descending from a common ancestor with the Chloroflexota. Quantitative analyses of gene duplication, transfer, origination, and loss across the bacterial tree reveal that HGT affects the majority of genes (92%), while 66% of genes show vertical transmission, making a tree a suitable framework on which to examine bacterial evolution. The reconciliations analysis also allows for mapping of gene families to specific nodes in the tree, which was the foundation for the reconstruction of the metabolism and structure of LBCA. Based on posterior probabilities of gene family presence, LBCA is proposed to have already been a complex free-living, rod-shaped diderm, with flagella, pili, and lipopolysaccharides. These, among other findings, suggest that LBCA was motile and capable of chemotaxis and environmental sensing. The recovery of components of glycolysis, the tricarboxylic-acid cycle (TCA), and the pentose-phosphate pathway provides evidence for carbohydrate metabolism. Carbon-fixation pathways were patchily resolved, with some evidence for the TCA, reductive glycine pathway, and the methyl branch of the WLP. While the presence of enzymes comprising the Rhodobacter nitrogen-fixation (Rnf) complex would be consistent with the possibility of acetogenic growth, the primary catalytic enzyme of the WLP was not recovered. A complete CRISPR-Cas system implies LBCA may have already been in contact with viruses and other parasitic replicators.

CHAPTER 5: A NEW VIEW OF THE TREE OF LIFE INCLUDING THE VIROSPHERE

Chapter 5 is a literature review addressing advances in microbial sampling, genome sequencing, and phylogenetics that have reshaped the TOL and how the recently proposed "Virosphere" contributes to the diversification of life and cellular evolution. The root of the infamous 3D tree sits between a branch leading to Bacteria and another branch leading to the sister clades, Archaea and Eukaryota. Over the past several decades, genomic and phylogenetic evidence have led to a revised scenario of eukaryogenesis, one that involved an ancestral archaealhost and one or more bacterial partners. The discovery of the Asgard archaea as the closest phylogenetic relatives of eukaryotes was a central element supporting the archaeal origins of eukaryotes, and the subsequent cultivation of the first *Lokiarchaeota* provided morphological and physiological evidence underlying possible mechanisms of eukaryogenesis. Robust computational techniques, such as outgroup free rooting methods and limiting phylogenetic artifacts have placed the CPR bacteria as a derived lineage sister to the Chloroflexota, whereas additional analyses are required to confirm the basal placement of their counterparts in the Archaea, the DPANN. Deeper sequencing of eukaryotic genomes has resolved additional branches in the eukaryotic tree, resulting in better resolution of the root of the eukaryotes and the nature of LECA. Viruses, other non-cellular parasites, and mobile genetic elements (MGEs) are generally excluded from the TOL, however they are

ubiquitous in the biosphere and have been shown to be instrumental in genome evolution and ecology. Genomics techniques have been useful in categorizing viruses and other MGEs into a tiered multi-realm sphere, termed the Virosphere. All but one of the six major realms of the virosphere are believed to have originated around or before LUCA, making them crucial in our understanding of cellular evolution from its earliest stages. The current proposal of viral evolution is one that combines two major branches that run parallel to the origins of life from the primordial environment. The replication (genetic) module is proposed to have originated from the primordial replicon pool, while the morphogenetic (encapsulation) module was acquired several times throughout cellular history, mirrored in the diversification of viral structure. Overcoming sequencing and technical limitations would expand progress on addressing debate surrounding life's origins and evolution, including the parallel involvement of the virosphere.

CHAPTER 6: ARCHAEAL DIVERSITY, TAXONOMY, AND ECOLOGY

Chapter 6 is a book chapter that takes a closer look at evolving perspectives on the Archaea, including their placement in the TOL, taxonomy, diversity, and their ecological roles. The diversity and evolution of the Archaea has undergone massive expansion in the past 50 years since the discovery of the domain. Initial observations revealed that Archaea shared many features with eukaryotes to the exclusion of Bacteria, inviting speculation over their evolutionary position in the TOL and their role in eukaryogenesis. Widespread sampling of diverse environmental samples coupled with cultivation-independent techniques have revealed the ubiquity of archaea across a wide variety of environments. Improved sequencing and phylogenetic methods have identified new branches in the archaeal tree highlighting the diversity of these organisms. The discovery of the Asgard superphylum indicated a branch within the Archaea that leads to the eukaryotes, with ancestral members of this group proposed to be the ancestral host of the protoeukaryote. In addition, the presence of a diverse and deeply-branching archaeal radiation known as the DPANN superphylum, has opened unanswered questions regarding the evolutionary processes at play in the archaeal tree. Representatives of the DPANN are generally reduced in cell and genome size, and based on the patchy presence of key biosynthetic pathways and other cellular features, are believed to primarily depend on symbiotic relationships with a more metabolically complete host. Genomic and experimental evidence has implicated Archaea as critical players in global nutrient cycling, for example the role of methanogens and methanotrophs in cycling methane in low-energy environments. In addition, emerging evidence has predicted members of this group to play an important role in the human microbiome, including oral, gut, and lung health.

REFERENCES

1. J. B. S. Haldane, The origin of life: Rationalist Annual., v. 148. (1929).

2. H. C. Urey, On the Early Chemical History of the Earth and the Origin of Life. *Proceedings of the National Academy of Sciences* **38**, 351–363 (1952).

3. M. N. Georgieva, C. T. S. Little, V. V. Maslennikov, A. G. Glover, N. R. Ayupova, R. J. Herrington, The history of life at hydrothermal vents. *Earth-Sci. Rev.* **217**, 103602 (2021).

4. J. B. Corliss, J. Dymond, L. I. Gordon, J. M. Edmond, R. P. von Herzen, R. D. Ballard, K. Green, D. Williams, A. Bainbridge, K. Crane, T. H. van Andel, Submarine thermal springs on the Galápagos Rift. *Science* **203**, 1073–1083 (1979).

5. D. S. Kelley, J. A. Karson, D. K. Blackman, G. L. Früh-Green, D. A. Butterfield, M. D. Lilley, E. J. Olson, M. O. Schrenk, K. K. Roe, G. T. Lebon, P. Rivizzigno, AT3-60 Shipboard Party, An off-axis hydrothermal vent field near the Mid-Atlantic Ridge at 30 degrees N. *Nature* **412**, 145–149 (2001).

6. A. Omran, M. Pasek, A Constructive Way to Think about Different Hydrothermal Environments for the Origins of Life. *Life* **10** (2020).

7. S. Q. Lang, W. J. Brazelton, Habitability of the marine serpentinite subsurface: a case study of the Lost City hydrothermal field. *Philos. Trans. A Math. Phys. Eng. Sci.* **378**, 20180429 (2020).

8. W. Martin, M. J. Russell, On the origin of biochemistry at an alkaline hydrothermal vent. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **362**, 1887–1925 (2007).

9. J. H. E. Cartwright, M. J. Russell, The origin of life: the submarine alkaline vent theory at 30. *Interface Focus* **9**, 20190104 (2019).

10. M. Paecht-Horowitz, Clays as possible catalysts for peptide formation in the prebiotic era. *Orig. Life* **7**, 369–381 (1976).

11. D. E. LaRowe, P. Regnier, Thermodynamic potential for the abiotic synthesis of adenine, cytosine, guanine, thymine, uracil, ribose, and deoxyribose in hydrothermal systems. *Orig. Life Evol. Biosph.* **38**, 383–397 (2008).

12. G. Proskurowski, M. D. Lilley, J. S. Seewald, G. L. Früh-Green, E. J. Olson, J. E. Lupton, S. P. Sylva, D. S. Kelley, Abiogenic hydrocarbon production at lost city hydrothermal field. *Science* **319**, 604–607 (2008).

13. M. J. Russell, W. Martin, The rocky roots of the acetyl-CoA pathway. *Trends Biochem. Sci.* **29**, 358–363 (2004).

14. S. W. Ragsdale, Enzymology of the wood-Ljungdahl pathway of acetogenesis. *Ann. N. Y. Acad. Sci.* **1125**, 129–136 (2008).

15. J. E. Goldford, H. Hartman, T. F. Smith, D. Segrè, Remnants of an Ancient Metabolism without Phosphate. *Cell* **168**, 1126-1134.e9 (2017).

16. G. Fuchs, Alternative pathways of carbon dioxide fixation: insights into the early evolution of life? *Annu. Rev. Microbiol.* **65**, 631–658 (2011).

17. P. Mitchell, Coupling of phosphorylation to electron and hydrogen transfer by a chemi-osmotic type of mechanism. *Nature* **191**, 144–148 (1961).

18. P. Mitchell, J. Moyle, Chemiosmotic hypothesis of oxidative phosphorylation. *Nature* **213**, 137–139 (1967).

19. G. F. Joyce, J. W. Szostak, Protocells and RNA Self-Replication. *Cold Spring Harb. Perspect. Biol.* **10** (2018).

20. J. P. Schrum, T. F. Zhu, J. W. Szostak, The origins of cellular life. *Cold Spring Harb. Perspect. Biol.* **2**, a002212 (2010).

21. E. V. Koonin, Carl Woese's vision of cellular evolution and the domains of life. *RNA Biol.* **11**, 197–204 (2014).

22. H. M. B. Harris, C. Hill, A Place for Viruses on the Tree of Life. *Front. Microbiol.* **11**, 604048 (2020).

23. C. R. Woese, G. E. Fox, The concept of cellular evolution. *J. Mol. Evol.* **10**, 1–6 (1977).

24. W. F. Doolittle, J. R. Brown, Tempo, mode, the progenote, and the universal root. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 6721–6728 (1994).

25. C. Woese, The universal ancestor. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 6854–6859 (1998).

26. D. Penny, A. Poole, The nature of the last universal common ancestor. *Curr. Opin. Genet. Dev.* **9**, 672–677 (1999).

27. W. F. Doolittle, The nature of the universal ancestor and the evolution of the proteome. *Curr. Opin. Struct. Biol.* **10**, 355–358 (2000).

28. P. Forterre, The origin of DNA genomes and DNA replication proteins. *Curr. Opin. Microbiol.* **5**, 525–532 (2002).

29. F. L. Sousa, T. Thiergart, G. Landan, S. Nelson-Sathi, I. A. C. Pereira, J. F. Allen, N. Lane, W. F. Martin, Early bioenergetic evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368**, 20130088 (2013).

30. B. Schoepp-Cothenet, R. van Lis, A. Atteia, F. Baymann, L. Capowiez, A.-L. Ducluzeau, S. Duval, F. ten Brink, M. J. Russell, W. Nitschke, On the universal core of bioenergetics. *Biochim. Biophys. Acta* **1827**, 79–93 (2013).

31. V. Sojo, A. Pomiankowski, N. Lane, A bioenergetic basis for membrane divergence in archaea and bacteria. *PLoS Biol.* **12**, e1001926 (2014).

32. M. C. Weiss, F. L. Sousa, N. Mrnjavac, S. Neukirchen, M. Roettger, S. Nelson-Sathi, W. F. Martin, The physiology and habitat of the last universal common ancestor. *Nat Microbiol* **1**, 16116 (2016).

33. M. J. Russell, W. Nitschke, Methane: Fuel or Exhaust at the Emergence of Life? *Astrobiology* **17**, 1053–1066 (2017).

34. M. C. Weiss, M. Preiner, J. C. Xavier, V. Zimorski, W. F. Martin, The last universal common ancestor between ancient Earth chemistry and the onset of genetics. *PLoS Genet.* **14**, e1007518 (2018).

35. D. D. Leipe, L. Aravind, E. V. Koonin, Did DNA replication evolve twice independently? *Nucleic Acids Res.* **27**, 3389–3401 (1999).

36. P. Forterre, J. Filée, H. Myllykallio, "Origin and Evolution of DNA and DNA Replication Machineries" in *The Genetic Code and the Origin of Life*, L. Ribas de Pouplana, Ed. (Springer US, Boston, MA, 2004), pp. 145–168.

37. P. Forterre, 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. U. S. A.* **103**, 3669–3674 (2006).

38. A. I. Oparin, *Proiskhozhdenie Zhizni* (Voennoe Izd. Ministerstva Obrony Sojuza SSR, 1924).

39. A. I. Oparin, The origin of life.[Transl. by S. Morgulis.]. *Macmillan; New York, NY* (1938).

40. S. L. Miller, J. W. Schopf, A. Lazcano, Oparin's ``Origin of Life'': Sixty Years Later. *J. Mol. Evol.* **44**, 351–353 (1997).

41. A. Lazcano, Historical development of origins research. *Cold Spring Harb. Perspect. Biol.* **2**, a002089 (2010).

42. J. Peretó, Out of fuzzy chemistry: from prebiotic chemistry to metabolic networks. *Chem. Soc. Rev.* **41**, 5394–5403 (2012).

43. P. Schönheit, W. Buckel, W. F. Martin, On the origin of heterotrophy. *Trends Microbiol.* **24**, 12–25 (2016).

44. W. Martin, M. J. Russell, On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **358**, 59–83; discussion 83-5 (2003).

45. G. Wächtershäuser, From volcanic origins of chemoautotrophic life to Bacteria, Archaea and Eukarya. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **361**, 1787–806; discussion 1806-8 (2006).

46. N. Lane, J. F. Allen, W. Martin, How did LUCA make a living? Chemiosmosis in the origin of life. *Bioessays* **32**, 271–280 (2010).

47. V. Sojo, B. Herschy, A. Whicher, E. Camprubí, N. Lane, The Origin of Life in Alkaline Hydrothermal Vents. *Astrobiology* **16**, 181–197 (2016).

48. P. S. Adam, G. Borrel, S. Gribaldo, Evolutionary history of carbon monoxide dehydrogenase/acetyl-CoA synthase, one of the oldest enzymatic complexes. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E1166–E1173 (2018).

49. C. von Linnaeus, Systema naturae, vol. 1. *Systema naturae, Vol. 1* (1758).

50. A. van Leewenhoeck, Observations, communicated to the publisher by Mr. Antony van Leewenhoeck, in a dutch letter of the 9th Octob. 1676. here English'd: concerning little animals by him observed in rain-well-sea- and snow water; as also in water wherein pepper had lain infused. *Philos. Trans. R. Soc. Lond.* **12**, 821–831 (1677).

51. C. Darwin, *On the Origin of Species by Means of Natural Selection* (Murray, London, 1859).

52. G. E. Budd, R. P. Mann, The dynamics of stem and crown groups. *Sci Adv* **6**, eaaz1626 (2020).

53. J. Sapp, The prokaryote-eukaryote dichotomy: Meanings and mythology. *Microbiol. Mol. Biol. Rev.* **69**, 292–305 (2005).

54. C. R. Woese, G. E. Fox, Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5088–5090 (1977).

55. C. R. Woese, O. Kandler, M. L. Wheelis, Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 4576–4579 (1990).

56. N. Iwabe, K. Kuma, M. Hasegawa, S. Osawa, T. Miyata, Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 9355–9359 (1989).

57. J. A. Lake, E. Henderson, M. Oakes, M. W. Clark, Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **81**, 3786–3790 (1984).

58. T. M. Embley, W. Martin, Eukaryotic evolution, changes and challenges. *Nature* **440**, 623–630 (2006).

59. T. A. Williams, T. M. Embley, Archaeal "dark matter" and the origin of eukaryotes. *Genome Biol. Evol.* **6**, 474–481 (2014).

60. T. A. Williams, T. M. Embley, Changing ideas about eukaryotic origins. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140318 (2015).

61. H. Noller, Carl Woese (1928-2012). *Nature* **493**, 610 (2013).

62. E. A. Eloe-Fadrosh, N. N. Ivanova, T. Woyke, N. C. Kyrpides, Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nat Microbiol* **1**, 15032 (2016).

63. N. C. Kyrpides, E. A. Eloe-Fadrosh, N. N. Ivanova, Microbiome Data Science: Understanding Our Microbial Planet. *Trends Microbiol.* **24**, 425–427 (2016).

64. D. H. Parks, C. Rinke, M. Chuvochina, P.-A. Chaumeil, B. J. Woodcroft, P. N. Evans, P. Hugenholtz, G. W. Tyson, Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* **2**, 1533–1542 (2017).

65. Y. Wang, Y. Zhao, A. Bollas, Y. Wang, K. F. Au, Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **39**, 1348–1365 (2021).

66. L. A. Hug, B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D. A. Relman, K. M. Finstad, R. Amundson, B. C. Thomas, J. F. Banfield, A new view of the tree of life. *Nat Microbiol* **1**, 16048 (2016).

67. C. T. Brown, L. A. Hug, B. C. Thomas, I. Sharon, C. J. Castelle, A. Singh, M. J. Wilkins, K. C. Wrighton, K. H. Williams, J. F. Banfield, Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).

68. C. Rinke, P. Schwientek, A. Sczyrba, N. N. Ivanova, I. J. Anderson, J.-F. Cheng, A. Darling, S. Malfatti, B. K. Swan, E. A. Gies, J. A. Dodsworth, B. P. Hedlund, G. Tsiamis, S. M. Sievert, W.-T. Liu, J. A. Eisen, S. J. Hallam, N. C. Kyrpides, R. Stepanauskas, E. M. Rubin, P. Hugenholtz, T. Woyke, Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).

69. C. J. Castelle, K. C. Wrighton, B. C. Thomas, L. A. Hug, C. T. Brown, M. J. Wilkins, K. R. Frischkorn, S. G. Tringe, A. Singh, L. M. Markillie, R. C. Taylor, K. H. Williams, J. F. Banfield, Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr. Biol.* **25**, 690–701 (2015).

70. C. J. Castelle, C. T. Brown, K. Anantharaman, A. J. Probst, R. H. Huang, J. F. Banfield, Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat. Rev. Microbiol.* **16**, 629–645 (2018).

71. C. J. Castelle, J. F. Banfield, Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. *Cell* **172**, 1181–1197 (2018).

72. U. Jahn, M. Gallenberger, W. Paper, B. Junglas, W. Eisenreich, K. O. Stetter, R. Rachel, H. Huber, Nanoarchaeum equitans and Ignicoccus hospitalis: new insights into a unique, intimate association of two archaea. *J. Bacteriol.* **190**, 1743–1750 (2008).

73. J. N. Hamm, S. Erdmann, E. A. Eloe-Fadrosh, A. Angeloni, L. Zhong, C. Brownlee, T. J. Williams, K. Barton, S. Carswell, M. A. Smith, S. Brazendale, A. M. Hancock, M. A. Allen, M. J. Raftery, R. Cavicchioli, Unexpected host dependency of Antarctic Nanohaloarchaeota. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 14661–14670 (2019).

74. E. St. John, Y. Liu, M. Podar, M. B. Stott, J. Meneghin, Z. Chen, K. Lagutin, K. Mitchell, A.-L. Reysenbach, A new symbiotic nanoarchaeote (Candidatus Nanoclepta minutus) and its host (Zestosphaera tikiterensis gen. nov., sp. nov.) from a New Zealand hot spring. *Syst. Appl. Microbiol.* **42**, 94–106 (2019).

75. X. He, J. S. McLean, A. Edlund, S. Yooseph, A. P. Hall, S.-Y. Liu, P. C. Dorrestein, E. Esquenazi, R. C. Hunter, G. Cheng, K. E. Nelson, R. Lux, W. Shi, Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 244–249 (2015).

76. B. Bor, J. S. McLean, K. R. Foster, L. Cen, T. T. To, A. Serrato-Guillen, F. E. Dewhirst, W. Shi, X. He, Rapid evolution of decreased host susceptibility drives a stable relationship between ultrasmall parasite TM7x and its bacterial host. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 12277–12282 (2018).

77. D. Moreira, Y. Zivanovic, A. I. López-Archilla, M. Iniesto, P. López-García, Reductive evolution and unique predatory mode in the CPR bacterium Vampirococcus lugosii. *Nat. Commun.* **12**, 2454 (2021).

78. N. Dombrowski, T. A. Williams, J. Sun, B. J. Woodcroft, J.-H. Lee, B. Q. Minh, C. Rinke, A. Spang, Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. *Nat. Commun.* **11**, 3939 (2020).

79. R. Méheust, D. Burstein, C. J. Castelle, J. F. Banfield, The distinction of CPR bacteria from other bacteria based on protein family content. *Nat. Commun.* **10**, 4173 (2019).

80. C. Brochier-Armanet, P. Forterre, S. Gribaldo, Phylogeny and evolution of the Archaea: one hundred genomes later. *Curr. Opin. Microbiol.* **14**, 274–281 (2011).

81. C. Petitjean, P. Deschamps, P. López-García, D. Moreira, Rooting the domain archaea by phylogenomic analysis supports the foundation of the new kingdom Proteoarchaeota. *Genome Biol. Evol.* **7**, 191–204 (2014).

82. M. Aouad, N. Taib, A. Oudart, M. Lecocq, M. Gouy, C. Brochier-Armanet, Extreme halophilic archaea derive from two distinct methanogen Class II lineages. *Mol. Phylogenet. Evol.* **127**, 46–54 (2018).

83. Y. Feng, U. Neri, S. Gosselin, A. S. Louyakis, R. T. Papke, U. Gophna, J. P. Gogarten, The Evolutionary Origins of Extreme Halophilic Archaeal Lineages. *Genome Biol. Evol.* **13** (2021).

84. N. A. Moran, G. M. Bennett, The tiniest tiny genomes. *Annu. Rev. Microbiol.* **68**, 195–215 (2014).

85. J. Bergsten, A review of long-branch attraction. *Cladistics* **21**, 163–193 (2005).

86. H. C. Betts, M. N. Puttick, J. W. Clark, T. A. Williams, P. C. J. Donoghue, D. Pisani, Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nat Ecol Evol* **2**, 1556–1562 (2018).

87. D. Chernikova, S. Motamedi, M. Csürös, E. V. Koonin, I. B. Rogozin, A late origin of the extant eukaryotic diversity: divergence time estimates using rare genomic changes. *Biol. Direct* **6**, 26 (2011).

88. L. Eme, S. C. Sharpe, M. W. Brown, A. J. Roger, On the age of eukaryotes: evaluating evidence from fossils and molecular clocks. *Cold Spring Harb. Perspect. Biol.* **6** (2014).

89. P. C. J. Donoghue, C. Kay, A. Spang, G. Szöllősi, A. Nenarokova, E. R. R. Moody, D. Pisani, T. A. Williams, Defining eukaryotes to dissect eukaryogenesis. *Curr. Biol.* **33**, R919–R929 (2023).

90. A. F. W. Schimper, Über die entwicklung der chlorophyllkörner und farbkörper. *Bot. Ztg.* **41**, 105 (1883).

91. W. Martin, K. V. Kowallik, Annotated English translation of Mereschkowsky's 1905 paper 'Über Natur und Ursprung der Chromatophoren im Pflanzenreiche.' *Eur. J. Phycol.* **34**, 287–295 (1999).

92. P. Portier, *Les Symbiotes* (Masson, 1918).

93. L. Sagan, On the origin of mitosing cells. *J. Theor. Biol.* **14**, 255–274 (1967).

94. A. Spang, J. H. Saw, S. L. Jørgensen, K. Zaremba-Niedzwiedzka, J. Martijn, A. E. Lind, R. van Eijk, C. Schleper, L. Guy, T. J. G. Ettema, Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).

95. K. Zaremba-Niedzwiedzka, E. F. Caceres, J. H. Saw, D. Bäckström, L. Juzokaite, E. Vancaester, K. W. Seitz, K. Anantharaman, P. Starnawski, K. U. Kjeldsen, M. B. Stott, T. Nunoura, J. F. Banfield, A. Schramm, B. J. Baker, A. Spang, T. J. G. Ettema, Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).

96. T. A. Williams, C. J. Cox, P. G. Foster, G. J. Szöllősi, T. M. Embley, Phylogenomics provides robust support for a two-domains tree of life. *Nat Ecol Evol* **4**, 138–147 (2020).

97. J. A. Lake, Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. *Nature* **331**, 184–186 (1988).

98. C. J. Cox, P. G. Foster, R. P. Hirt, S. R. Harris, T. M. Embley, The archaebacterial origin of eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 20356–20361 (2008).

99. P. G. Foster, C. J. Cox, T. M. Embley, The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 2197–2207 (2009).

100. L. Guy, T. J. G. Ettema, The archaeal "TACK" superphylum and the origin of eukaryotes. *Trends Microbiol.* **19**, 580–587 (2011).

101. T. A. Williams, P. G. Foster, T. M. W. Nye, C. J. Cox, T. M. Embley, A congruent phylogenomic signal places eukaryotes within the Archaea. *Proc. Biol. Sci.* **279**, 4870–4879 (2012).

102. T. A. Williams, P. G. Foster, C. J. Cox, T. M. Embley, An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* **504**, 231–236 (2013).

103. L. Guy, J. H. Saw, T. J. G. Ettema, The archaeal legacy of eukaryotes: a phylogenomic perspective. *Cold Spring Harb. Perspect. Biol.* **6**, a016022 (2014).

104. K. Raymann, C. Brochier-Armanet, S. Gribaldo, The two-domain tree of life is linked to a new root for the Archaea. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 6670–6675 (2015).

105. C. Rinke, M. Chuvochina, A. J. Mussig, P.-A. Chaumeil, A. A. Davín, D. W. Waite, W. B. Whitman, D. H. Parks, P. Hugenholtz, A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nat Microbiol* **6**, 946–959 (2021).

106. Y. Liu, K. S. Makarova, W.-C. Huang, Y. I. Wolf, A. N. Nikolskaya, X. Zhang, M. Cai, C.-J. Zhang, W. Xu, Z. Luo, L. Cheng, E. V. Koonin, M. Li, Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature* **593**, 553–557 (2021).

107. L. Eme, D. Tamarit, E. F. Caceres, C. W. Stairs, V. De Anda, M. E. Schön, K. W. Seitz, N. Dombrowski, W. H. Lewis, F. Homa, J. H. Saw, J. Lombard, T. Nunoura, W.-J. Li, Z.-S. Hua, L.-X. Chen, J. F. Banfield, E. S. John, A.-L. Reysenbach, M. B. Stott, A. Schramm, K. U. Kjeldsen, A. P. Teske, B. J. Baker, T. J. G. Ettema, Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature* **618**, 992–999 (2023).

108. K. W. Seitz, C. S. Lazar, K.-U. Hinrichs, A. P. Teske, B. J. Baker, Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. *ISME J.* **10**, 1696–1705 (2016).

109. L. E. Valentin-Alvarado, K. E. Appler, V. De Anda, M. C. Schoelmerich, J. West-Roberts, V. Kivenson, A. Crits-Christoph, L. Ly, R. Sachdeva, D. F. Savage, B. J. Baker, J. F. Banfield, Asgard archaea modulate potential methanogenesis substrates in wetland soil, *bioRxiv* (2023)p. 2023.11.21.568159.

110. C. Akıl, S. Ali, L. T. Tran, J. Gaillard, W. Li, K. Hayashida, M. Hirose, T. Kato, A. Oshima, K. Fujishima, L. Blanchoin, A. Narita, R. C. Robinson, Structure and dynamics of Odinarchaeota tubulin and the implications for eukaryotic microtubule evolution. *Sci Adv* **8**, eabm2225 (2022).

111. B. Henneman, C. van Emmerik, H. van Ingen, R. T. Dame, Structure and function of archaeal histones. *PLoS Genet.* **14**, e1007582 (2018).

112. H. Hartman, A. Fedorov, The origin of the eukaryotic cell: a genomic investigation. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 1420–1425 (2002).

113. F. Wu, D. R. Speth, A. Philosof, A. Crémière, A. Narayanan, R. A. Barco, S. A. Connon, J. P. Amend, I. A. Antoshechkin, V. J. Orphan, Unique mobile elements and scalable gene flow at the prokaryote–eukaryote boundary revealed by circularized Asgard archaea genomes. *Nature Microbiology* **7**, 200–212 (2022).

114. T. Hatano, S. Palani, D. Papatziamou, R. Salzer, D. P. Souza, D. Tamarit, M. Makwana, A. Potter, A. Haig, W. Xu, D. Townsend, D. Rochester, D. Bellini, H. M. A. Hussain, T. J. G. Ettema, J. Löwe, B. Baum, N. P. Robinson, M. Balasubramanian, Asgard archaea shed light on the evolutionary origins of the eukaryotic ubiquitin-ESCRT machinery. *Nat. Commun.* **13**, 3398 (2022).

115. T. Rodrigues-Oliveira, F. Wollweber, R. I. Ponce-Toledo, J. Xu, S. K.-M. R. Rittmann, A. Klingl, M. Pilhofer, C. Schleper, Actin cytoskeleton and complex cell architecture in an Asgard archaeon. *Nature* **613**, 332–339 (2023).

116. V. Da Cunha, M. Gaia, D. Gadelle, A. Nasir, P. Forterre, Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS Genet.* **13**, e1006810 (2017).

117. V. Da Cunha, M. Gaia, A. Nasir, P. Forterre, Asgard archaea do not close the debate about the universal tree of life topology, *PLoS genetics*. **14** (2018)p. e1007215.

118. A. Spang, L. Eme, J. H. Saw, E. F. Caceres, K. Zaremba-Niedzwiedzka, J. Lombard, L. Guy, T. J. G. Ettema, Asgard archaea are the closest prokaryotic relatives of eukaryotes, *PLoS genetics*. **14** (2018)p. e1007080.

119. S. G. Garg, N. Kapust, W. Lin, M. Knopp, F. D. K. Tria, S. Nelson-Sathi, S. B. Gould, L. Fan, R. Zhu, C. Zhang, W. F. Martin, Anomalous Phylogenetic Behavior of Ribosomal Proteins in Metagenome-Assembled Asgard Archaea. *Genome Biol. Evol.* **13** (2021).

120. H. Imachi, M. K. Nobu, N. Nakahara, Y. Morono, M. Ogawara, Y. Takaki, Y. Takano, K. Uematsu, T. Ikuta, M. Ito, Y. Matsui, M. Miyazaki, K. Murata, Y. Saito, S. Sakai, C. Song, E. Tasumi, Y. Yamanaka, T. Yamaguchi, Y. Kamagata, H. Tamaki, K. Takai, Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**, 519–525 (2020).

121. D. Tamarit, E. F. Caceres, M. Krupovic, R. Nijland, L. Eme, N. P. Robinson, T. J. G. Ettema, A closed Candidatus Odinarchaeum chromosome exposes Asgard archaeal viruses. *Nat Microbiol* **7**, 948–952 (2022).

122. J. Vosseberg, J. J. E. van Hooff, M. Marcet-Houben, A. van Vlimmeren, L. M. van Wijk, T. Gabaldón, B. Snel, Timing the origin of eukaryotic cellular complexity with ancient duplications. *Nat Ecol Evol* **5**, 92–100 (2021).

123. D. Yang, Y. Oyaizu, H. Oyaizu, G. J. Olsen, C. R. Woese, Mitochondrial origins. *Proc. Natl. Acad. Sci. U. S. A.* **82**, 4443–4447 (1985).

124. A. J. Roger, S. A. Muñoz-Gómez, R. Kamikawa, The Origin and Diversification of Mitochondria. *Curr. Biol.* **27**, R1177–R1192 (2017).

125. S. G. Andersson, A. Zomorodipour, J. O. Andersson, T. Sicheritz-Pontén, U. C. Alsmark, R. M. Podowski, A. K. Näslund, A. S. Eriksson, H. H. Winkler, C. G. Kurland, The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature* **396**, 133–140 (1998).

126. D. A. Fitzpatrick, C. J. Creevey, J. O. McInerney, Genome phylogenies indicate a meaningful alpha-proteobacterial phylogeny and support a grouping of the mitochondria with the Rickettsiales. *Mol. Biol. Evol.* **23**, 74–85 (2006).

127. Z. Wang, M. Wu, An integrated phylogenomic approach toward pinpointing the origin of mitochondria. *Sci. Rep.* **5**, 7949 (2015).

128. T. Gabaldón, Relative timing of mitochondrial endosymbiosis and the "pre-mitochondrial symbioses" hypothesis. *IUBMB Life* **70**, 1188–1196 (2018).

129. L. Fan, D. Wu, V. Goremykin, J. Xiao, Y. Xu, S. Garg, C. Zhang, W. F. Martin, R. Zhu, Phylogenetic analyses with systematic taxon sampling show that mitochondria branch within Alphaproteobacteria. *Nat Ecol Evol* **4**, 1213–1219 (2020).

130. J. Martijn, J. Vosseberg, L. Guy, P. Offre, T. J. G. Ettema, Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* **557**, 101–105 (2018).

131. S. A. Muñoz-Gómez, E. Susko, K. Williamson, L. Eme, C. H. Slamovits, D. Moreira, P. López-García, A. J. Roger, Site-and-branch-heterogeneous analyses of an expanded dataset favour mitochondria as sister to known Alphaproteobacteria. *Nat Ecol Evol* **6**, 253–262 (2022).

132. J. Martijn, J. Vosseberg, L. Guy, P. Offre, T. J. G. Ettema, Phylogenetic affiliation of mitochondria with Alpha-II and Rickettsiales is an artefact, *Nature ecology & evolution*. **6** (2022)pp. 1829–1831.

133. J. F. H. Strassert, I. Irisarri, T. A. Williams, F. Burki, A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. *Nat. Commun.* **12**, 1879 (2021).

134. S. J. Sibbald, J. M. Archibald, Genomic Insights into Plastid Evolution. *Genome Biol. Evol.* **12**, 978–990 (2020).

135. R. I. Ponce-Toledo, P. Deschamps, P. López-García, Y. Zivanovic, K. Benzerara, D. Moreira, An Early-Branching Freshwater Cyanobacterium at the Origin of Plastids. *Curr. Biol.* **27**, 386–391 (2017).

136. A. M. Poole, S. Gribaldo, Eukaryotic origins: How and when was the mitochondrion acquired? *Cold Spring Harb. Perspect. Biol.* **6**, a015990 (2014).

137. E. V. Koonin, Origin of eukaryotes from within archaea, archaeal eukaryome and bursts of gene gain: eukaryogenesis just made easier? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140333 (2015).

138. P. López-García, D. Moreira, The Syntrophy hypothesis for the origin of eukaryotes revisited. *Nat Microbiol* **5**, 655–667 (2020).

139. T. Gabaldón, Origin and Early Evolution of the Eukaryotic Cell. *Annu. Rev. Microbiol.* **75**, 631–647 (2021).

140. W. Martin, M. Müller, The hydrogen hypothesis for the first eukaryote. *Nature* **392**, 37–41 (1998).

141. P. López-Garćia, D. Moreira, Metabolic symbiosis at the origin of eukaryotes. *Trends Biochem. Sci.* **24**, 88–93 (1999).

142. A. Spang, C. W. Stairs, N. Dombrowski, L. Eme, J. Lombard, E. F. Caceres, C. Greening, B. J. Baker, T. J. G. Ettema, Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. *Nat Microbiol* **4**, 1138–1148 (2019).

143. J. Martijn, T. J. G. Ettema, From archaeon to eukaryote: the evolutionary dark ages of the eukaryotic cell. *Biochem. Soc. Trans.* **41**, 451–457 (2013).

144. T. J. G. Ettema, Evolution: Mitochondria in the second act, *Nature*. **531** (2016)pp. 39–40.

145. A. A. Pittis, T. Gabaldón, Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature* **531**, 101–104 (2016).

146. G. Mendel, Versuche uber pflanzen-hybriden. *Vorgelegt in den Sitzungen* (1865).

147. D. J. Fairbanks, S. Abbott, Darwin's Influence on Mendel: Evidence from a New Translation of Mendel's Paper. *Genetics* **204**, 401–405 (2016).

148. P. Portin, A. Wilkins, The Evolving Definition of the Term "Gene." *Genetics* **205**, 1353–1364 (2017).

149. L. Loewe, W. G. Hill, The population genetics of mutations: good, bad and indifferent. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**, 1153–1167 (2010).

150. M. Kimura, The neutral theory of molecular evolution. *Sci. Am.* **241**, 98–100, 102, 108 passim (1979).

151. M. Kimura, The neutral theory of molecular evolution: a review of recent evidence. *Jpn. J. Genet.* **66**, 367–386 (1991).

152. E. V. Koonin, Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* **39**, 309–338 (2005).

153. E. V. Koonin, Towards a postmodern synthesis of evolutionary biology. *Cell Cycle* **8**, 799–800 (2009).

154. J. P. Gogarten, H. Kibak, P. Dittrich, L. Taiz, E. J. Bowman, B. J. Bowman, M. F. Manolson, R. J. Poole, T. Date, T. Oshima, J. Konishi, K. Denda, M. Yoshida, Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 6661–6665 (1989).

155. J. P. Gogarten, L. Taiz, Evolution of proton pumping ATPases: Rooting the tree of life. *Photosynth. Res.* **33**, 137–146 (1992).

156. F. Griffith, The Significance of Pneumococcal Types. *J. Hyg.* **27**, 113–159 (1928).

157. H. Ochman, J. G. Lawrence, E. A. Groisman, Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).

158. E. V. Koonin, K. S. Makarova, L. Aravind, Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.* **55**, 709–742 (2001).

159. P. J. Keeling, J. D. Palmer, Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.* **9**, 605–618 (2008).

160. S. S. Abby, E. Tannier, M. Gouy, V. Daubin, Lateral gene transfer as a support for the tree of life. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 4962–4967 (2012).

161. V. Daubin, G. J. Szöllősi, Horizontal Gene Transfer and the History of Life. *Cold Spring Harb. Perspect. Biol.* **8**, a018036 (2016).

162. K. B. Sieber, R. E. Bromley, J. C. Dunning Hotopp, Lateral gene transfer between prokaryotes and eukaryotes. *Exp. Cell Res.* **358**, 421–426 (2017).

163. C. W. Stairs, J. E. Dharamshi, D. Tamarit, L. Eme, S. L. Jørgensen, A. Spang, T. J. G. Ettema, Chlamydial contribution to anaerobic metabolism during eukaryotic evolution. *Sci Adv* **6**, eabb7258 (2020).

164. S. J. Sibbald, L. Eme, J. M. Archibald, A. J. Roger, Lateral Gene Transfer Mechanisms and Pan-genomes in Eukaryotes. *Trends Parasitol.* **36**, 927–941 (2020).

165. T. Gabaldón, E. V. Koonin, Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.* **14**, 360–366 (2013).

166. T. Dagan, W. Martin, The tree of one percent. *Genome Biol.* **7**, 118 (2006).

167. Q. Zhu, U. Mai, W. Pfeiffer, S. Janssen, F. Asnicar, J. G. Sanders, P. Belda-Ferre, G. A. Al-Ghalith, E. Kopylova, D. McDonald, T. Kosciolek, J. B. Yin, S. Huang, N. Salam, J.-Y. Jiao, Z. Wu, Z. Z. Xu, K. Cantrell, Y. Yang, E. Sayyari, M. Rabiee, J. T. Morton, S. Podell, D. Knights, W.-J. Li, C. Huttenhower, N. Segata, L. Smarr, S. Mirarab, R. Knight, Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477 (2019).

168. J. H. Degnan, N. A. Rosenberg, Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* **24**, 332–340 (2009).

169. O. R. P. Bininda-Emonds, J. L. Gittleman, M. A. Steel, The (Super)Tree of Life: Procedures, Problems, and Prospects. *Annu. Rev. Ecol. Syst.* **33**, 265–289 (2002).

170. O. R. P. Bininda-Emonds, The evolution of supertrees. *Trends Ecol. Evol.* **19**, 315–322 (2004).

171. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

172. A. Krogh, M. Brown, I. S. Mian, K. Sjölander, D. Haussler, Hidden Markov models in computational biology. *J. Mol. Biol.* **235**, 1501–1531 (1994).

173. S. R. Eddy, A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**, 205–211 (2009).

174. S. R. Eddy, Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).

175. D. M. Portik, J. J. Wiens, Do Alignment and Trimming Methods Matter for Phylogenomic (UCE) Analyses? *Syst. Biol.* **70**, 440–462 (2021).

176. J. Felsenstein, Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376 (1981).

177. J. P. Huelsenbeck, F. Ronquist, MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).

178. P. Lemey, M. Salemi, A.-M. Vandamme, *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing* (Cambridge University Press, 2009).

179. F. F. Nascimento, M. D. Reis, Z. Yang, A biologist's guide to Bayesian phylogenetic analysis. *Nat Ecol Evol* **1**, 1446–1454 (2017).

180. C. Kosiol, L. Bofkin, S. Whelan, Phylogenetics by likelihood: evolutionary modeling as a tool for understanding the genome. *J. Biomed. Inform.* **39**, 51–61 (2006).

181. T. H. Jukes, C. R. Cantor, "Evolution of Protein Molecules" in *Mammalian Protein Metabolism* (Elsevier, 1969), pp. 21–132.

182. S. Tavaré, Some probabilistic and statistical problems in the analysis of DNA sequences. *Lecture of Mathematics for Life Science* (1986).

183. S. Q. Le, O. Gascuel, An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).

184. Z. Yang, Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306–314 (1994).

185. L. S. Quang, O. Gascuel, N. Lartillot, Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **24**, 2317–2323 (2008).

186. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

187. J. Felsenstein, CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP. *Evolution* **39**, 783–791 (1985).

188. D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

189. F. Lemoine, J.-B. Domelevo Entfellner, E. Wilkinson, D. Correia, M. Dávila Felipe, T. De Oliveira, O. Gascuel, Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* **556**, 452–456 (2018).

190. D. M. Hillis, J. J. Bull, An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. *Syst. Biol.* **42**, 182–192 (1993).

191. H. Kishino, M. Hasegawa, Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J. Mol. Evol.* **29**, 170–179 (1989).

192. H. Shimodaira, M. Hasegawa, Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114–1116 (1999).

193. H. Shimodaira, An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508 (2002).

194. G. J. Szöllősi, W. Rosikiewicz, B. Boussau, E. Tannier, V. Daubin, Efficient exploration of the space of reconciled gene trees. *Syst. Biol.* **62**, 901–912 (2013).

195. T. A. Williams, G. J. Szöllősi, A. Spang, P. G. Foster, S. E. Heaps, B. Boussau, T. J. G. Ettema, T. M. Embley, Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E4602–E4611 (2017).

196. H. Sauquet, A practical guide to molecular dating. *C. R. Palevol* **12**, 355–367 (2013).

197. E. Zuckerkandl, Molecular disease, evolution, and genic heterogeneity. *Horiz. Biochem. Biophys.*, 189–225 (1962).

198. G. J. Szöllősi, S. Höhna, T. A. Williams, D. Schrempf, V. Daubin, B. Boussau, Relative Time Constraints Improve Molecular Dating. *Syst. Biol.* **71**, 797–809 (2022).

199. A. A. Davín, E. Tannier, T. A. Williams, B. Boussau, V. Daubin, G. J. Szöllősi, Gene transfers can date the tree of life. *Nat Ecol Evol* **2**, 904–909 (2018).

200. P. M. Shih, N. J. Matzke, Primary endosymbiosis events date to the later Proterozoic with cross-calibrated phylogenetic dating of duplicated ATPase proteins. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 12355–12360 (2013).

201. P. P. Sharma, W. C. Wheeler, Cross-bracing uncalibrated nodes in molecular dating improves congruence of fossil and molecular age estimates. *Front. Zool.* **11**, 1–13 (2014).

202. K. Kruger, P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling, T. R. Cech, Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell* **31**, 147–157 (1982).

203. C. Petibon, M. Malik Ghulam, M. Catala, S. Abou Elela, Regulation of ribosomal protein genes: An ordered anarchy. *Wiley Interdiscip. Rev. RNA* **12**, e1632 (2021).

204. A. G. Stewart, E. M. Laming, M. Sobti, D. Stock, Rotary ATPases--dynamic molecular machines. *Curr. Opin. Struct. Biol.* **25**, 40–48 (2014).

205. A. Y. Mulkidjanian, K. S. Makarova, M. Y. Galperin, E. V. Koonin, Inventing the dynamo machine: the evolution of the F-type and V-type ATPases. *Nat. Rev. Microbiol.* **5**, 892–899 (2007).

206. A. Y. Mulkidjanian, M. Y. Galperin, K. S. Makarova, Y. I. Wolf, E. V. Koonin, Evolutionary primacy of sodium bioenergetics. *Biol. Direct* **3**, 13 (2008).

207. R. L. Cross, V. Müller, The evolution of A-, F-, and V-type ATP synthases and ATPases: reversals in function and changes in the H+/ATP coupling ratio. *FEBS Lett.* **576**, 1–4 (2004).

208. A. Y. Mulkidjanian, P. Dibrov, M. Y. Galperin, The past and present of sodium energetics: may the sodium-motive force be with you. *Biochim. Biophys. Acta* **1777**, 985–992 (2008).

209. K. Schlegel, V. Leone, J. D. Faraldo-Gómez, V. Müller, Promiscuous archaeal ATP synthase concurrently coupled to Na+ and H+ translocation. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 947–952 (2012).

210. G. Grüber, M. S. S. Manimekalai, F. Mayer, V. Müller, ATP synthases from archaea: the beauty of a molecular motor. *Biochim. Biophys. Acta* **1837**, 940–952 (2014).

211. S. Mirarab, R. Reaz, M. S. Bayzid, T. Zimmermann, M. S. Swenson, T. Warnow, ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541-8 (2014).

CHAPTER 2

An estimate of the deepest branches of the tree of life from ancient vertically evolving genes

Edmund R.R. Moody, Tara A. Mahendrarajah, Nina Dombrowski, James W. Clark, Celine Petitjean, Pierre Offre, Gergely J. Szöllősi, Anja Spang, Tom A. Williams

eLife, 2022

SUMMARY AND CONTRIBUTIONS

Marker gene selection can have a profound impact on the topology of a phylogeny. A recent analysis (*1*) inferred a tree of 10,575 Archaea and Bacteria using 381 marker genes (hereafter, expanded set) using a combination of a supertree and conventional gene concatenation approach and resolved a short interdomain branch (hereafter, AB branch), which challenges traditional findings of a distant relationship between the two primary domains. Concatenated phylogenetic inference is traditionally implemented using a small subset of vertically evolving genes. Despite their efficacy in resolving phylogenetic relationships, small sets of genes only represent a minute fraction of the totality of genomic content. Consequently, including a more diverse set of marker genes can provide additional genetic information from which to inform evolutionary relationships. Zhu and coworkers used the supertree method, ASTRAL (*2*, *3*), to infer a tree of Archaea and Bacteria, and traditional concatenated phylogenetic inference to compute branch lengths, which were then applied to the supertree topology. The disparity between the results obtained by Zhu et. al. versus more traditional approaches prompted us to investigate and compare the various approaches to assess their reliability (e.g. utility of concatenated phylogenetic inference methods, marker gene selection, model fit, and trimming methods). Specifically, we combined manual inspection of marker gene trees, quantification of the taxonomic splits between established groups, phylogenetic inference, and model testing to assess how well different marker gene sets 1) resolve the deepest events in the tree of life (TOL) and 2) can be used to determine divergence time between the primary domains of life. When inspecting the marker gene phylogenies from the published study (*1*), we found that the majority of markers exhibited evolutionary histories that were incongruent with the vertically evolving processes occurring in the TOL. Namely, we found most markers failed to meet key criteria, such as reciprocal monophyly of the two domains and limited horizontal gene transfer (HGT). Our inspection revealed that hidden paralogy and interdomain HGT contributed the most to artificially shortening the AB branch. We observed similar trends in an in-house curated set of marker genes from previously published studies (*4*–*6*), which we used to statistically determine the best performing markers and infer a phylogeny that recovered a long AB branch and placement of relevant clades. This suggests that the observation of a shorter interdomain branch, as presented by Zhu and coworkers, is a result of phylogenetic artifacts rather than reflective of deep cellular evolution.

This project was a collaborative effort that required major contributions from authors with expertise in different fields. Active analysis and writing for this project occurred over the course of approximately 24 months. We held collaborative weekly writing and analysis meetings for approximately 13 months until the paper was published. My contribution to this project included conceptualization, data curation, formal analysis and investigation, as well as writing of the original draft, revision, and final draft. My data analysis involved the following major sections: the inspection, evaluation, and testing of the 381 marker genes used by Zhu and co workers to infer a phylogeny of 10,575 Archaea and Bacteria (*1*), 2) the inspection, evaluation, and testing of a curated set of marker genes from previously published analyses (*4*–*6*) to assess

reliability and conditions necessary for phylogenetic inference, 3) taxonomic selection for our focal analysis, and 4) taxonomic evaluation of the 10,575 Archaea and Bacteria used by Zhu et. al. (*1*). I manually inspected each single gene tree in the expanded dataset to identify what proportion of genes recovered the reciprocal monophyly of the primary domains, as well as for the presence of paralogous gene families (i.e., based on sequence annotations I performed on all sequences in the published dataset). The manual assessment of the extent of domain paraphyly, HGT, and presence of paralogs was summarized and used to determine reliable marker gene selection methods. This analysis was coupled with a statistical approach that quantified the number of times established taxonomic relationships were recovered in trees generated using the expanded set of marker genes (but with a subset of taxa for computational tractability). Briefly, the number of "splits" is correlated to the number of times a particular lineage is recovered monophyletic; it is defined by a metric known as the *split-score* (*7*). I summarized the split-score statistics to identify which genes were the "highest-ranking" markers based on the quantified ranking process. I also performed a taxonomy check using the GTDB-Toolkit in an effort to properly establish taxonomic groups for all taxa including in the initial study (*1*), which I used for the split-score analysis and the visualization of the distribution of taxonomic groups compared to our working dataset.

The second part of this project involved curation of a collection of 95 ribosomal, non-ribosomal, and core marker genes (*4–6*) from previously published studies to examine the attributes and performance of different marker genes in phylogenetic reconstruction. Likewise, I annotated all sequences from the 350 Archaea and 350 Bacteria in our dataset and collected orthologs corresponding to all 95 markers. The KO, Pfam, and their corresponding descriptions were mapped to the tips of the trees and used during several rounds of manual inspection. I applied the split-score ranking procedure to the 54 markers that met reciprocal monophyly of Archaea and Bacteria, and the downstream statistical analysis recovered 27 marker genes that were the 50% best performing markers from the dataset. The resulting phylogeny of those markers revealed a long AB branch length and diversification of the major domains that is consistent with recent findings.

References

1. Q. Zhu, U. Mai, W. Pfeiffer, S. Janssen, F. Asnicar, J. G. Sanders, P. Belda-Ferre, G. A. Al-Ghalith, E. Kopylova, D. McDonald, T. Kosciolek, J. B. Yin, S. Huang, N. Salam, J.-Y. Jiao, Z. Wu, Z. Z. Xu, K. Cantrell, Y. Yang, E. Sayyari, M. Rabiee, J. T. Morton, S. Podell, D. Knights, W.-J. Li, C. Huttenhower, N. Segata, L. Smarr, S. Mirarab, R. Knight, Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477 (2019).
2. S. Mirarab, R. Reaz, M. S. Bayzid, T. Zimmermann, M. S. Swenson, T. Warnow, ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541-8 (2014).
3. C. Zhang, M. Rabiee, E. Sayyari, S. Mirarab, ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**, 153 (2018).
4. C. Petitjean, P. Deschamps, P. López-García, D. Moreira, Rooting the domain archaea by phylogenomic analysis supports the foundation of the new kingdom Proteoarchaeota. *Genome Biol. Evol.* **7**, 191–204 (2014).
5. T. A. Williams, C. J. Cox, P. G. Foster, G. J. Szöllősi, T. M. Embley, Phylogenomics provides robust support for a two-domains tree of life. *Nat Ecol Evol* **4**, 138–147 (2020).
6. G. A. Coleman, A. A. Davín, T. A. Mahendrarajah, L. L. Szánthó, A. Spang, P. Hugenholtz, G. J. Szöllősi, T. A. Williams, A rooted phylogeny resolves early bacterial evolution. *Science* **372** (2021).
7. N. Dombrowski, T. A. Williams, J. Sun, B. J. Woodcroft, J.-H. Lee, B. Q. Minh, C. Rinke, A. Spang, Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. *Nat. Commun.* **11**, 3939 (2020).

ABSTRACT

Core gene phylogenies provide a window into early evolution, but different gene sets and analytical methods have yielded substantially different views of the tree of life. Trees inferred from a small set of universal core genes have typically supported a long branch separating the archaeal and bacterial domains. By contrast, recent analyses of a broader set of non-ribosomal genes have suggested that Archaea may be less divergent from Bacteria, and that estimates of inter-domain distance are inflated due to accelerated evolution of ribosomal proteins along the inter-domain branch. Resolving this debate is key to determining the diversity of the archaeal and bacterial domains, the shape of the tree of life, and our understanding of the early course of cellular evolution. Here, we investigate the evolutionary history of the marker genes key to the debate. We show that estimates of a reduced Archaea-Bacteria (AB) branch length result from inter-domain gene transfers and hidden paralogy in the expanded marker gene set. By contrast, analysis of a broad range of manually curated marker gene datasets from an evenly sampled set of 700 Archaea and Bacteria reveals that current methods likely underestimate the AB branch length due to substitutional saturation and poor model fit; that the best-performing phylogenetic markers tend to support longer inter-domain branch lengths; and that the AB branch lengths of ribosomal and non-ribosomal marker genes are statistically indistinguishable. Furthermore, our phylogeny inferred from the 27 highest-ranked marker genes recovers a clade of DPANN at the base of the Archaea and places the bacterial Candidate Phyla Radiation (CPR) within Bacteria as the sister group to the Chloroflexota.

INTRODUCTION

Much remains unknown about the earliest period of cellular evolution and the deepest divergences in the tree of life. Phylogenies encompassing both Archaea and Bacteria have been inferred from a 'universal core' set of 16–56 genes encoding proteins involved in translation and other aspects of the genetic information processing machinery (*1–11*). While representing a small fraction of the total genome of any organism (*12*), these genes are thought to predominantly evolve vertically and are thus best suited for reconstructing the tree of life (*1*, *7*, *9*, *13*, *14*). In these analyses, the branch separating Archaea from Bacteria (hereafter, the AB branch) is often the longest internal branch in the tree (*4*, *10*, *15–18*). In molecular phylogenetics, branch lengths are usually measured in expected numbers of substitutions per site, with a long branch corresponding to a greater degree of genetic change. Long branches can therefore result from high evolutionary rates, long periods of absolute time, or a combination of the two. If a sufficient number of fossils are available for calibration, molecular clock models can, in principle, disentangle the contributions of these effects. However, limited fossil data (*19*) is currently available to calibrate early divergences in the tree of life (*20–23*), and as a result, the ages and evolutionary rates of the deepest branches of the tree remain highly uncertain.

Recently, Zhu et al., 2019 (*24*) inferred a phylogeny from 381 genes distributed across Archaea and Bacteria using the supertree method ASTRAL (*25*). These markers increase the total number of genes compared to other universal marker sets and comprise not only proteins involved in information processing but also proteins affiliated with most other functional COG categories, including metabolic processes (Supplementary file 1). The genetic distance (AB branch length) between the domains (*24*) was estimated from a concatenation of the same marker genes, resulting in a much shorter AB branch length than observed with the core universal markers (*4*, *10*). These analyses were consistent with the hypothesis (*6*, *24*) that the apparent deep divergence of Archaea and Bacteria might be the result of an accelerated evolutionary rate of genes encoding translational and in particular ribosomal proteins along the AB branch as compared to other genes. Interestingly, the same observation was made previously using a smaller set of 38 non-ribosomal marker proteins (*6*), although the difference in AB branch length between ribosomal and non-ribosomal markers in that analysis was reported to be substantially lower (roughly twofold, compared to roughly 10-fold for the 381 protein set (*6*, *24*)).

A higher evolutionary rate of ribosomal genes might result from the accumulation of compensatory substitutions at the interaction surfaces among the protein subunits of the ribosome (*6*, *26*) or as a compensatory response to the addition or removal of ribosomal subunits early in evolution (*6*). Alternatively, differences in the inferred AB branch length might result from varying rates or patterns of evolution between the traditional core genes (*10*, *27*) and the expanded set (*24*). Substitutional saturation (multiple substitutions at the same site) (*28*) and across-site compositional heterogeneity can both impact the inference of tree topologies and branch lengths (*29–34*) These difficulties are particularly significant for

ancient divergences (*35*). Failure to model site-specific amino acid preferences has previously been shown to lead to underestimation of the AB branch length due to a failure to detect convergent changes (*10*, *36*), although the published analysis of the 381 marker set did not find evidence of a substantial impact of these features on the tree as a whole (*24*). Those analyses also identified phylogenetic incongruence among the 381 markers, but did not determine the underlying cause (*24*).

This recent work (Zhu et al., 2019) (*24*) raises two important issues regarding the inference of the universal tree: first, that estimates of the genetic distance between Archaea and Bacteria from classic 'core genes' may not be representative of ancient genomes as a whole, and second, that there may be many more suitable genes to investigate early evolutionary history than generally recognized, providing an opportunity to improve the precision and accuracy of deep phylogenies. Here, we investigate these issues in order to determine how different methodologies and marker sets affect estimates of the evolutionary distance between Archaea and Bacteria. First, we examine the evolutionary history of the 381-gene marker set (hereafter, the expanded marker gene set) and identify several features of these genes, including instances of inter-domain gene transfers and mixed paralogy, that may contribute to the inference of a shorter AB branch length in concatenation analyses. Then, we re-evaluate the marker gene sets used in a range of previous analyses to determine how these and other factors, including substitutional saturation and model fit, contribute to inter-domain branch length estimations and the shape of the universal tree. Finally, we identify a subset of marker genes least affected by these issues and use these to estimate an updated tree of the primary domains of life and the length of the branch that separates Archaea and Bacteria.

RESULTS AND DISCUSSION

GENES FROM THE EXPANDED MARKER SET ARE NOT WIDELY DISTRIBUTED IN ARCHAEA

The 381-gene set was derived from a larger set of 400 genes used to estimate the phylogenetic placement of new lineages as part of the PhyloPhlAn method (*37*) and applied to a taxonomic selection that included 669 Archaea and 9906 Bacteria (*24*). Perhaps reflecting the focus on Bacteria in the original application, the phylogenetic distribution of the 381 marker genes in the expanded set varies substantially (Supplementary file 1), with many being poorly represented in Archaea. Specifically, 41% of the published gene trees (https://biocore.github.io/wol/; Zhu et al., 2019 (*24*)) contain less than 25% of the sampled archaea, with 14 and 68 of these trees including 0 or ≤10 archaeal homologues, respectively. Across all of the gene trees, archaeal homologues comprise 0–14.8% of the dataset (Supplementary file 1). Manual inspection of subsampled versions of these gene trees suggested that 317/381 did not possess an unambiguous branch separating the archaeal and bacterial domains (Supplementary file 1). These distributions suggest that many of these genes are not broadly present in both domains, and that some might be specific to Bacteria.

CONFLICTING EVOLUTIONARY HISTORIES OF INDIVIDUAL MARKER GENES AND THE INFERRED SPECIES TREE

In the published analysis of the 381-gene set (*24*), the tree topology was inferred using the supertree method ASTRAL (*25*), with branch lengths inferred on this fixed tree from a marker gene concatenation (*24*). The topology inferred from this expanded marker set (*24*) is similar to previous trees (*4*, *38*) and recovers Archaea and Bacteria as reciprocally monophyletic domains, albeit with a shorter AB branch than in earlier analyses. However, the individual gene trees (*24*) differ regarding domain monophyly: Archaea and Bacteria are recovered as reciprocally monophyletic groups in only 22 of the 381 published (*24*) maximum likelihood (ML) gene trees of the expanded marker set (Supplementary file 1).

Since single-gene trees often fail to strongly resolve ancient relationships, we used approximately unbiased (AU) tests (*39*) to evaluate whether the failure to recover domain monophyly in the published ML trees is statistically supported. For computational tractability, we performed these analyses on a 1000-species subsample of the full 10,575-species dataset that was compiled in the original study (*24*). For 79 of the 381 genes, we could not perform the test because the gene family did not contain any archaeal homologues (56 genes) or contained only one archaeal homologue (23 genes); in total, the 1000-species sample included 74 archaeal genomes. For the remaining 302 genes, domain monophyly was rejected at the 5% significance level (with Bonferroni correction, $p<0.0001656$) for 151 out of 302 (50%) genes. As a comparison, we performed the same test on several smaller marker sets used previously to infer a tree of life (*6*, *10*, *40*); none of the markers in those sets rejected reciprocal domain monophyly ($p<0.05$ for all genes, with Bonferroni correction: Coleman: >0.001724; Petitjean: >0.001316; Williams: >0.00102: Fig. 1A). In what follows, we refer to four published marker gene sets as (i) the expanded set (381 genes; Zhu et al., 2019 (*24*)); (ii) the core set (49 genes; Williams et al., 2020 (*10*)), encoding ribosomal proteins and other conserved information-processing functions; itself a consensus set of several earlier studies (*27*, *41*, *42*); (iii) the non-ribosomal set (38 genes, broadly distributed and explicitly selected to avoid genes encoding ribosomal proteins; Petitjean et al., 2014 (*6*)); and (iv) the bacterial set (29 genes used in a recent analysis of bacterial phylogeny; Coleman et al., 2021 (*40*)).

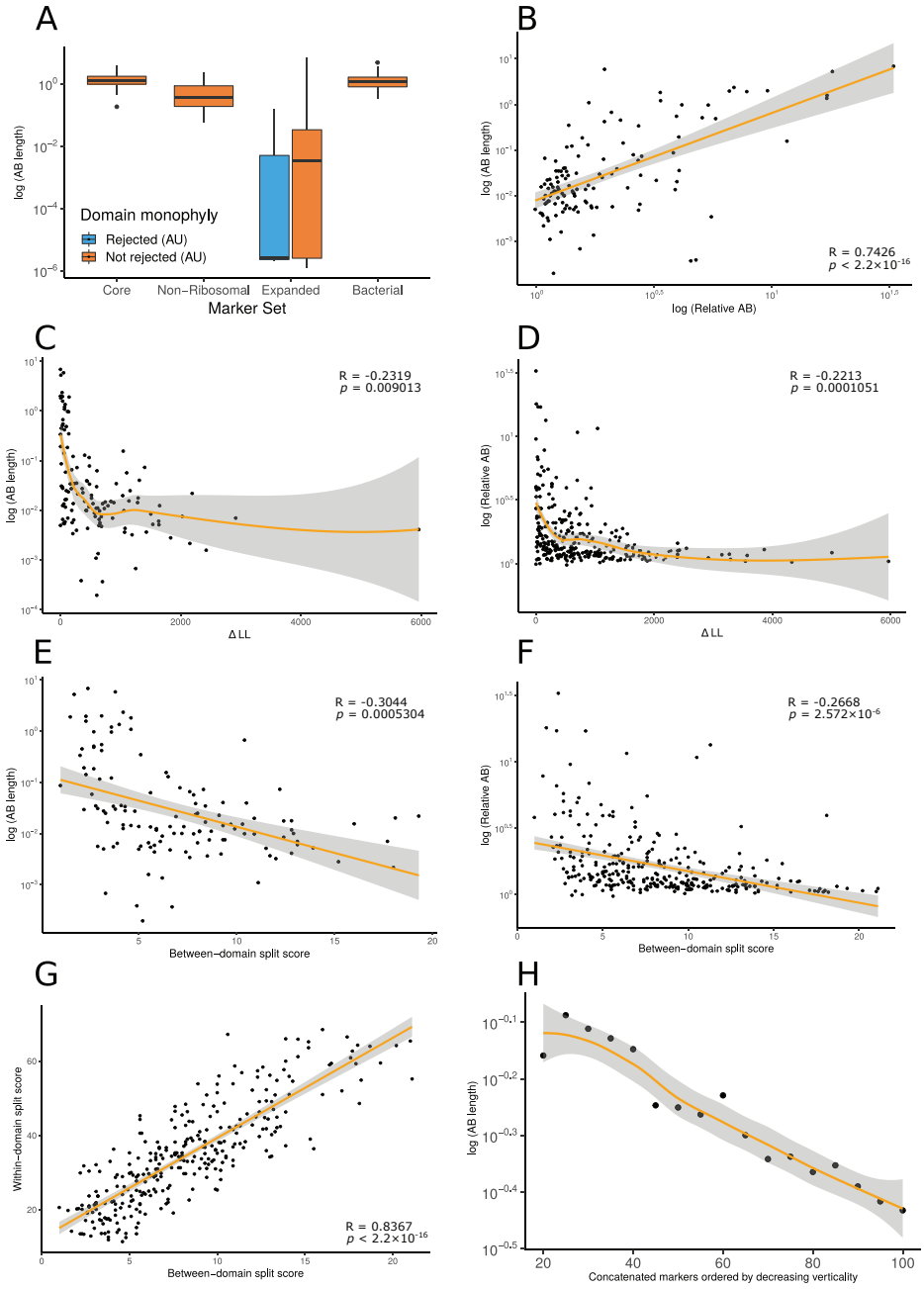


Fig. 1 | Vertically evolving marker genes support a greater evolutionary distance between Archaea and Bacteria. (**A**) Expanded set genes that reject domain monophyly (p<0.05, approximately unbiased [AU] test, with Bonferroni correction [see main text]) support significantly shorter Archaea-Bacteria (AB) branch lengths when constrained to follow a domain monophyletic tree (p=3.653 × 10^{-6}, Wilcoxon rank-sum test). None of the marker genes from several other published analyses significantly reject domain monophyly (Bonferroni-corrected p<0.05, AU test) for all genes tested, consistent with vertical inheritance

from the LUCA (last universal common ancestor) to the last common ancestors of Archaea and Bacteria, respectively. (**B**) Two measures of evolutionary proximity (*24*), AB branch length and relative AB distance, are positively correlated ($R = 0.7426499$, $p < 2.2 \times 10^{-16}$). We considered two complementary proxies of marker gene verticality: ΔLL (**C**: against AB branch length; **D**: against relative AB length), which reflects the degree to which marker genes reject domain monophyly (**C**: p=0.009013 and $R = -0.2317894$; **D**: p=0.0001051 and $R = -0.2213292$); and the between-domain split score (**E**: against AB branch length; **F**: against relative AB length), which quantifies the extent to which marker genes recover monophyletic Archaea and Bacteria; a higher split score (see Materials and methods) indicates the splitting of domains into multiple gene tree clades due to gene transfer, reciprocal sorting out of paralogs, or lack of phylogenetic resolution (**E**: p=0.0005304 and $R = -0.3043537$; **F**: $p = 2.572 \times 10^{-6}$ and $R = -0.2667739$). We also considered a split score based on within-domain relationships (**G**); between- and within-domain split scores are positively correlated: $R = 0.836679$, $p < 2.2 \times 10^{-16}$, Pearson's correlation, indicating that markers that recover Archaea and Bacteria as monophyletic also tend to recover established within-domain relationships. (**H**) Inferred AB length decreases as marker genes of lower verticality (larger ΔLL) are added to the concatenate. Marker genes were sorted by ΔLL, the difference in log-likelihood between the maximum likelihood gene family tree under a free topology search and the log-likelihood of the best tree constrained to obey domain monophyly. Note that 79/381 expanded set markers had zero or one archaea in the 1000-species subsample and so could not be included in these analyses; of the remaining 302 markers, 176 have AB branch lengths very close to 0 in the constraint tree (as seen in panel **A**). In these plots, we removed all markers with an AB branch length of <0.00001; see Fig. 1—figure supplements 1–13 for all plots. Nonlinear trendlines were estimated using LOESS regression.

Fig. supplement 1. No evidence for a relationship between Archaea-Bacteria (AB) branch length and gene evolutionary rate (average MAD root-to- tip distance).
Fig. supplement 2. No evidence for a relationship between Archaea-Bacteria (AB) branch length and gene evolutionary rate (average MAD root-to- tip distance).
Fig. supplement 3. A significant positive relationship between relative Archaea-Bacteria (AB) distance and evolutionary rate (MAD root-to-tip distance).
Fig. supplement 4. Two proxies for marker gene verticality, ΔLL and between-domain split score, are highly correlated.
Fig. supplement 5. Low-verticality genes (as measured by ΔLL) have a higher evolutionary rate (as measured by the mean root-to-tip distance on MAD-rooted gene trees).
Fig. supplement 6. Low-verticality genes (measured by between-domain split score) have a higher evolutionary rate (MAD root-to-tip distance).
Fig. supplement 7. High-verticality marker genes have longer Archaea-Bacteria (AB) branch lengths.
Fig. supplement 8. Archaea-Bacteria (AB) branch length and relative AB distance are positively correlated.
Fig. supplement 9. Archaea-Bacteria (AB) branch length is negatively correlated with between-domain split score.
Fig. supplement 10. Within-domain split score and ΔLL are strongly correlated, suggesting that both proxies capture a common signal of marker gene verticality.
Fig. supplement 11. Low-verticality marker genes (measured as within-domain split score) have shorter relative Archaea-Bacteria (AB) distances.
Fig. supplement 12. Low-verticality marker genes (measured as within-domain split score) have shorter Archaea-Bacteria (AB) branch lengths.
Fig. supplement 13. Low-verticality marker genes (measured as within-domain split score) have shorter Archaea-Bacteria (AB) branch lengths.
Fig. supplement 14. Raw count and percentage distribution of the Genome Taxonomy Database (GTDB)-defined classes for 10,575 archaeal and bacterial genomes in the expanded marker set analysis.
Fig. supplement 15. Raw count and percentage distribution of the Genome Taxonomy Database (GTDB)-defined phyla for 10,575 archaeal and bacterial genomes in the expanded marker set analysis.
Fig. supplement 16. Raw count and percentage distribution of domains for 10,575 archaeal and bacterial genomes in the expanded marker set analysis.

To investigate why 151 of the marker genes rejected the reciprocal monophyly of Archaea and Bacteria, we returned to the full dataset (*24*), annotated each sequence in each marker gene family by assigning proteins to KOs, PFAMs and Interpro domains, among others (Supplementary file 1, see Materials and methods for details), and manually inspected the tree topologies (Supplementary file 1). This revealed that the major cause of domain polyphyly observed in gene trees was inter-domain gene transfer (in 359 out of 381 gene trees [94.2%]) and mixing of sequences from distinct paralogous families (in 246 out of 381 gene trees [64.6%]). For instance, marker genes encoding ABC-type transporters (p0131, p0151, p0159, p0174, p0181, p0287, p0306, p0364), tRNA synthetases (i.e., p0000, p0011, p0020, p0091, p0094, p0202), and aminotransferases and dehydratases (i.e., p0073/4-aminobutyrate aminotransferase; p0093/3-isopropylmalate dehydratase) often comprised a mixture of paralogs.

Together, these analyses indicate that the evolutionary histories of the individual markers of the expanded set differ from each other and from the species tree. The original study investigated and acknowledged (*24*) the varying levels of congruence between the marker phylogenies and the species tree, but did not investigate the underlying causes. Our analyses establish the basis for these disagreements in terms of gene transfers and the mixing of orthologs and paralogs within and between domains. The estimation of genetic distance based on concatenation relies on the assumption that all of the genes in the supermatrix evolve on the same underlying tree; genes with different gene tree topologies violate this assumption and should not be concatenated because the topological differences among sites are not modeled, and so the impact on inferred branch lengths is difficult to predict. In practice, it is often difficult to be certain that all of the markers in a concatenate share the same gene tree topology, and the analysis proceeds on the hypothesis that a small proportion of discordant genes are not expected to seriously impact the inferred tree. However, the concatenated tree inferred from the expanded marker set differs from previous trees in that the genetic distance between Bacteria and Archaea is greatly reduced, such that the AB branch length appears comparable to distances among bacterial phyla (*24*). Since an accurate estimate of the AB branch length has a major bearing on unanswered questions regarding the root of the universal tree (*35*), we next evaluated the impact of the conflicting gene histories within the expanded marker set on inferred AB branch length.

THE INFERRED BRANCH LENGTH BETWEEN ARCHAEA AND BACTERIA IS SHORTENED BY INTER-DOMAIN GENE TRANSFER AND HIDDEN PARALOGY

To investigate the impact of gene transfers and mixed paralogy on the AB branch length inferred by gene concatenations (*24*), we compared branch lengths estimated from markers on the basis of whether or not they rejected domain monophyly in the expanded marker set (Fig. 1A). To estimate AB branch lengths for genes in which the domains were not monophyletic in the ML tree, we first performed a constrained ML search to find the best gene tree that was consistent with domain monophyly for each family under the LG + G4 + F model in IQ-TREE 2 (*43*). While it may seem strained to estimate the length of a branch that does not appear

in the ML tree, we reasoned that this approach would provide insight into the contribution of these genes to the AB branch length in the concatenation, in which they conflict with the overall topology. AB branch lengths were significantly ($p=3.653 \times 10^{-6}$, Wilcoxon rank-sum test) shorter for markers that rejected domain monophyly (Bonferroni-corrected $p<0.0001656$; Fig. 1A): the mean AB branch length was 0.00668 substitutions/site for markers that significantly rejected domain monophyly and 0.287 substitutions/site for markers that did not reject domain monophyly. This behavior might result from marker gene transfers reducing the number of fixed differences between the domains, so that the AB branch length in a tree in which Archaea and Bacteria are constrained to be reciprocally monophyletic will tend towards 0 as the number of transfers increases.

To test the hypothesis that phylogenetic incongruence among markers might reduce the inferred AB distance, we evaluated the relationship between AB distance and two complementary metrics of marker gene verticality: ΔLL, the difference in log-likelihood between the constrained ML tree and the ML gene tree (a proxy for the extent to which a marker gene rejects the reciprocal monophyly of Bacteria and Archaea), and the 'split score' (*44*), which measures the extent to which marker genes recover established relationships for defined taxonomic levels of interest (e.g., at the level of domain, phylum, or order), averaging over bootstrap distributions of gene trees to account for phylogenetic uncertainty (see Materials and methods). We evaluated split scores at both the between-domain and within-domain (Fig. 1—figure supplements 1–13) levels. ΔLL and between-domain split score were positively correlated with each other (Fig. 1—figure supplement 4) and negatively correlated with both AB stem length (Fig. 1C and E) and relative AB distance (Fig. 1D and F), an alternative metric (*24*) that compares average tip-to-tip distances within and between domains. Interestingly, between-domain and within-domain split scores were strongly positively correlated (Fig. 1G), and the same relationships between within-domain split score, AB branch length, and relative AB distance were observed (Fig. 1—figure supplements 11 and 12). Overall, these results suggest that genes that recover the reciprocal monophyly of Archaea and Bacteria also evolve more vertically within each domain, and that these vertically evolving marker genes support a longer AB branch and a greater AB distance. Indeed, Zhu et al., 2019 (*24*) also recovered a significant positive relationship between gene verticality and relative AB distance (see their Fig. 5E). Consistent with these inferences, AB branch lengths estimated using concatenation decreased as increasing numbers of low-verticality markers (i.e., markers with higher ΔLL) were added to the concatenate (Fig. 1). These results suggest that inter-domain gene transfers reduce the overall AB branch length when included in a concatenation.

An alternative explanation for the positive relationship between marker gene verticality and AB branch length could be that vertically evolving genes experience higher rates of sequence evolution. For a set of genes that originate at the same point on the species tree, the mean root-to-tip distance (measured in substitutions per site, for gene trees rooted using the MAD (minimal ancestor deviation) method; ref. (*45*)) provides a proxy of evolutionary rate. Mean root-to-tip distances were significantly positively correlated with ΔLL and between-domain

split score (ΔLL: $R = 0.1397803$, p=0.01506, split score: $R = 0.1705415$, p=0.002947; Fig. 1—figure supplements 5 and 6), indicating that vertically evolving genes evolve relatively slowly (note that large values of ΔLL and split score denote low verticality). Thus, the longer AB branches of vertically evolving genes do not appear to result from a faster evolutionary rate for these genes. Taken together, these results indicate that the inclusion of genes that do not support the reciprocal monophyly of Archaea and Bacteria, or their constituent taxonomic ranks, in the universal concatenate explains the reduced estimated AB branch length.

FINDING ANCIENT VERTICALLY EVOLVING GENES

To estimate the AB branch length and the phylogeny of prokaryotes using a dataset that resolves some of the issues identified above, we performed a meta-analysis of several previous studies to identify a consensus set of vertically evolving marker genes. We identified unique markers from these analyses by reference to the COG ontology (Supplementary file 2, ref. (*44*); ref. (*46*)), extracted homologous sequences from a representative sample of 350 archaeal and 350 bacterial genomes (Supplementary file 3), and performed iterative phylogenetics and manual curation to obtain a set of 54 markers that recovered archaeal and bacterial monophyly (see Materials and methods). Prior to manual curation, non-ribosomal markers had a greater number of HGTs (horizontal gene transfer) and cases of mixed paralogy. In particular, for the original set of 95 unique COG families (see 'Phylogenetic analyses' in Materials and methods), we rejected 41 families based on the inferred ML trees either due to a large degree of HGT, paralogous gene families, or LBA (long branch attraction). For the remaining 54 markers, the ML trees contained evidence of occasional recent HGT events. Strict monophyly was violated in 69% of the non-ribosomal and 29% of the ribosomal families. We manually removed the individual sequences that violated domain monophyly before realignment, trimming, and subsequent tree inference (see Materials and methods). These results imply that manual curation of marker genes is important for deep phylogenetic analyses, particularly when using non-ribosomal markers. Comparison of within-domain split scores for these 54 markers (Supplementary file 4) indicated that markers that better resolved established relationships within each domain also supported a longer AB branch length (Fig. 2A). Further, the AB branch length inferred from a concatenation of the 54 marker genes increased moderately following pruning of recent HGTs, from 1.734 substitutions/site (non-pruned) to 1.945 substitutions/site after manual pruning, consistent with the hypothesis that non-modeled inter-domain HGTs reduce the overall estimate of AB branch length when included in concatenations.

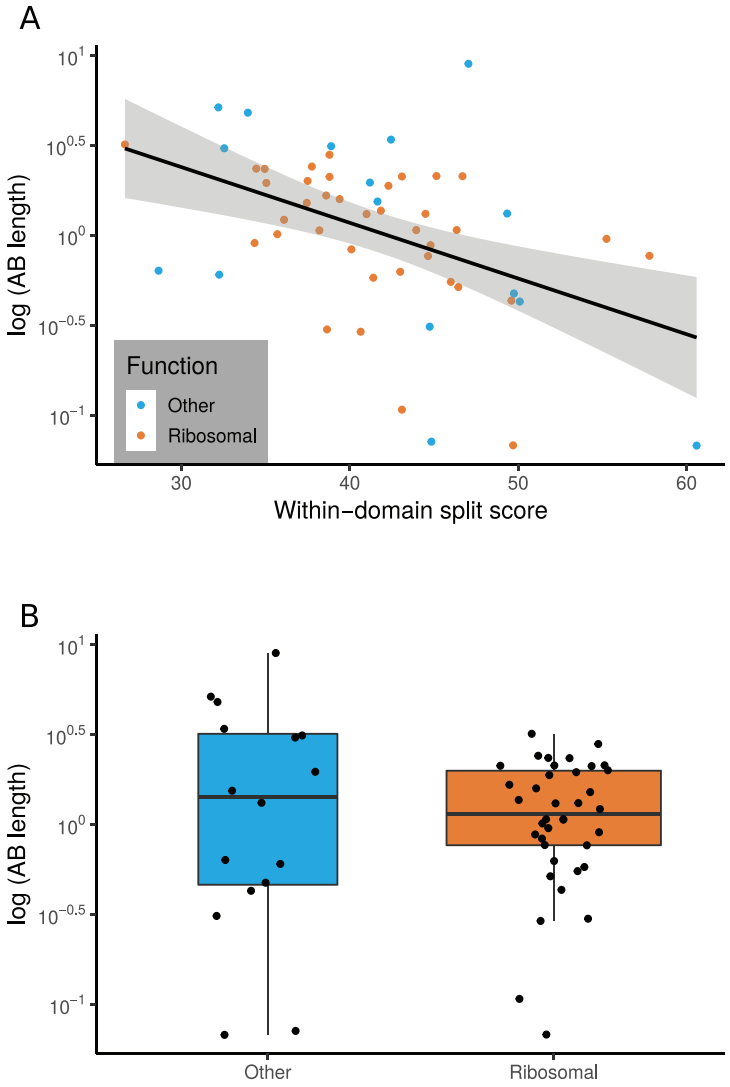


Fig. 2 | The relationship between marker gene verticality, Archaea-Bacteria (AB) branch length, and functional category. (**A**) Vertically evolving phylogenetic markers have longer AB branches. The plot shows the relationship between a proxy for marker gene verticality, within-domain split score (a lower split score denotes better recovery of established within-domain relationships, see Materials and methods), and AB branch length (in expected number of substitutions/site) for the 54 marker genes. Marker genes with higher split scores (that split established monophyletic groups into multiple subclades) have shorter AB branch lengths (p=0.0311, $R = 0.294$). Split scores of ribosomal and non-ribosomal markers were statistically indistinguishable (p=0.828, Fig. 2—figure supplement 1). (**B**) Among vertically evolving marker genes, ribosomal genes do not have a longer AB branch length. The plot shows functional classification of markers against AB branch length using 54 vertically evolving markers. We did not obtain a significant difference between AB branch lengths for ribosomal and non-ribosomal genes (p=0.6191, Wilcoxon rank-sum test).

Fig. supplement 1. Among vertically evolving marker genes, the split scores of ribosomal and non-ribosomal proteins are statistically indistinguishable.

DISTRIBUTIONS OF AB BRANCH LENGTHS FOR RIBOSOMAL AND NON-RIBOSOMAL MARKER GENES ARE SIMILAR

Traditional universal marker sets include many ribosomal proteins (*1–4*, *10*, *47*). If ribosomal proteins experienced accelerated evolution during the divergence of Archaea and Bacteria, this might lead to the inference of an artifactually long AB branch length (*6*, *24*). To investigate this, we plotted the inter-domain branch lengths for the 38 and 16 ribosomal and non-ribosomal genes, respectively, comprising the 54 marker genes set. We found no evidence that there was a longer AB branch associated with ribosomal markers than for other vertically evolving 'core' genes (Fig. 2B; mean AB branch length for ribosomal proteins 1.35 substitutions/site, mean for non-ribosomal 2.25 substitutions/site). To investigate further, we compared AB branch lengths inferred from concatenates of the ribosomal and non-ribosomal subsets of the 54 ancient, vertically evolving genes (Table 1). AB branch lengths from the ribosomal and non-ribosomal concatenates were similar overall, with some support for a longer AB branch length from vertically evolving non-ribosomal genes. Thus, these data do not support an accelerated evolutionary rate for ribosomal genes compared to other kinds of genes on the AB branch.

	AB branch length		Total tree length		AB branch length as a proportion of total tree length	
	Ribosomal	Non-ribosomal	Ribosomal	Non-ribosomal	Ribosomal	Non-ribosomal
27 marker set	1.9541	3.7723	250.7255	239.8203	0.0078	0.0157
54 marker set	1.8647	2.5414	271.3327	288.8470	0.0069	0.0088

Table 1
Archaea-Bacteria (AB) branch lengths and AB branch lengths as a proportion of total tree length inferred from ribosomal and non-ribosomal concatenates are similar.
The data do not support a faster evolutionary rate for ribosomal proteins on the AB branch compared to other kinds of ancient proteins.

SUBSTITUTIONAL SATURATION AND POOR MODEL FIT CONTRIBUTE TO UNDERESTIMATION OF AB BRANCH LENGTH

For the 27 most vertically evolving genes as ranked by within-domain split score, we performed an additional round of single-gene tree inference and manual review to identify and remove the remaining sequences that had evidence of HGT or represented distant paralogs. The resulting single-gene trees are provided in the Data Supplement (https://doi.org/10.6084/m9.figshare.13395470). To evaluate the relationship between site evolutionary rate and AB branch length, we created two concatenations: fastest sites (comprising sites with the highest probability of being in the fastest gamma rate category; 868 sites) and slowest sites (sites with the highest probability of being in the slowest gamma rate category, 1604 sites) and compared relative branch lengths inferred from the entire concatenate using IQ-TREE 2 to infer site-specific rates (Fig. 3).

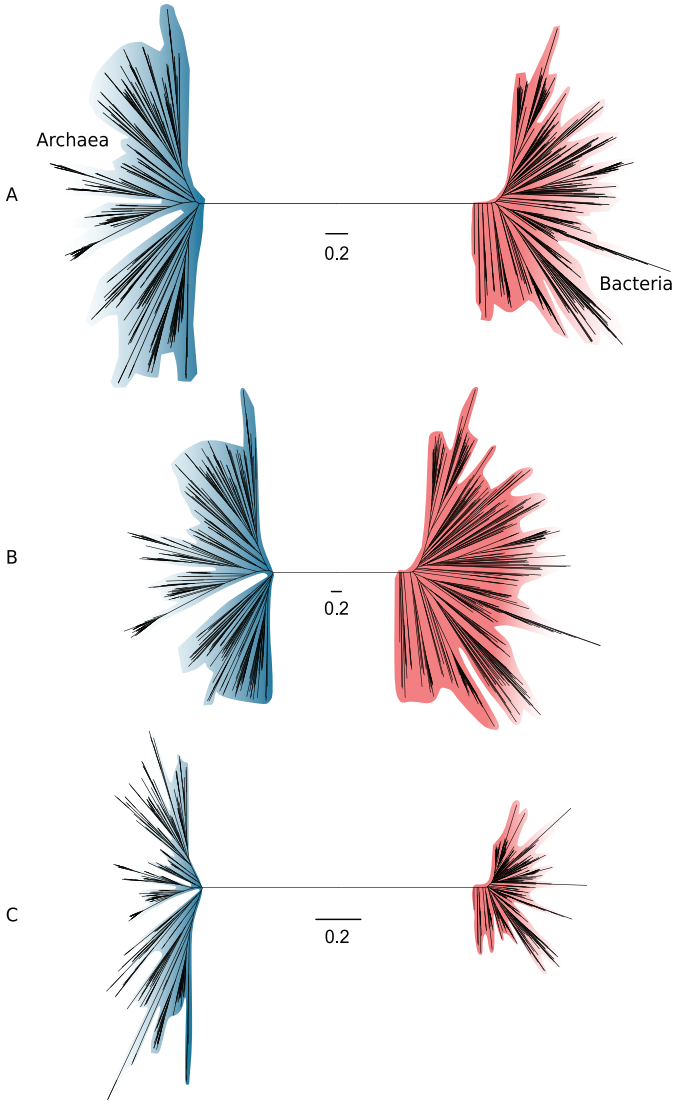


Fig. 3 | Slow- and fast-evolving sites support different shapes for the universal tree. (**A**) Tree of Archaea (blue) and Bacteria (red) inferred from a concatenation of 27 core genes using the best-fitting model (LG + C60 + G4 + F). (**B**) Tree inferred from the fastest-evolving sites. (**C**) Tree inferred from the slowest-evolving sites. To facilitate comparison of relative diversity, scale bars are provided separately for each panel; for a version of this figure with a common scale bar for all three panels, see Fig. 3—figure supplement 1. Slow-evolving sites support a relatively long inter-domain branch and less diversity within the domains (i.e., shorter between-taxa branch lengths within domains). This suggests that substitution saturation (overwriting of earlier changes) may reduce the relative length of the AB branch at fast-evolving sites and genes.

Fig. supplement 1. Slow- and fast-evolving sites support different shapes for the universal tree.
Fig. supplement 2. Vertically evolving genes and slow-evolving sites support a longer relative Archaea-Bacteria (AB) branch length.
Fig. supplement 3. The effect of modeling site compositional heterogeneity on Archaea-Bacteria (AB) branch length.

Notably, the proportion of inferred substitutions that occur along the AB branch differs between the slow-evolving and fast-evolving sites. As would be expected, the total tree length measured in substitutions per site is shorter from the slow-evolving sites, but the relative AB branch length is longer (1.2 substitutions/site, or ~2% of all inferred substitutions, compared to 2.6 substitutions/site, or ~0.04% of all inferred substitutions for the fastest-evolving sites; see Fig. 3—figure supplement 1 for absolute tree size comparisons). Since we would not expect the distribution of substitutions over the tree to differ between slow-evolving and fast-evolving sites, this result suggests that some ancient changes along the AB branch at fast-evolving sites have been overwritten by more recent events in evolution – that is, that substitutional saturation leads to an underestimate of the AB branch length (this is the case for both the expanded marker set and the 27 marker set; Fig. 3—figure supplement 2).

Another factor that has been shown to lead to underestimation of genetic distance on deep branches is a failure to adequately model the site-specific features of sequence evolution (*10*, *24*, *31*, *48*, *49*). Amino acid preferences vary across the sites of a sequence alignment due to variation in the underlying functional constraints (*31–33*). The consequence is that, at many alignment sites, only a subset of the 20 possible amino acids are tolerated by selection. Standard substitution models such as LG + G4 + F are site-homogeneous and approximate the composition of all sites using the average composition across the entire alignment. Such models underestimate the rate of evolution at highly constrained sites because they do not account for the high number of multiple substitutions that occur at such sites. The effect is that site-homogeneous models underestimate branch lengths when fit to site-heterogeneous data. Site-heterogeneous models have been developed that account for site-specific amino acid preferences, and these generally show improved fit to real protein sequence data (reviewed in ref. (*34*)). To evaluate the impact of substitution models on estimates of AB branch length, we assessed the fit of a range of models to the full concatenation using the Bayesian information criterion (BIC) in IQ-TREE 2. The AB branch length inferred under the best-fit model, the site-heterogeneous LG + C60 + G4 + F model, was 2.52 substitutions/site, ~1.7-fold greater than the branch length inferred from the site-homogeneous LG + G4 + F model (1.45 substitutions/site). Thus, substitution model fit has a major effect on the estimated length of the AB branch, with better-fitting models supporting a longer branch length (Table 2). The same trends are evident when better-fitting site-heterogeneous models are used to analyze the expanded marker set: considering only the top 5% of genes by ΔLL score, the AB branch length is 1.2 under LG + G4 + F, but increases to 2.4 under the best-fitting LG + C60 + G4 + F model (Fig. 3—figure supplement 3). These results are consistent with Zhu et al., 2019 (*24*), who also noted that AB branch length increases as model fit improves for the expanded marker dataset.

Overall, these results indicate that difficulties with modeling sequence evolution, either due to substitutional saturation or failure to model variation in site compositions, lead to an underestimation of the AB branch length, both in published analyses and for the analyses of the new dataset presented here. As substitution models improve, we would therefore expect estimates of the AB branch length to increase further.

Substitution model	BIC (ΔBIC)	AB branch length
LG + G4 + F	5935950.053	1.4491
LG + C20 + G4 + F	(152046.1)	2.1394
LG + C40 + G4 + F	(179126.7)	2.4697
LG + C60 + G4 + F	(189063.8)	2.5178

Table 2
The inferred Archaea-Bacteria (AB) branch length from a concatenation of the top 27 markers using a simple model compared to models that account for site compositional heterogeneity.
Models that account for across-site compositional heterogeneity fit the data better (as assessed by lower Bayesian information criterion [BIC] scores) and infer a longer AB branch length.

A PHYLOGENY OF ARCHAEA AND BACTERIA INFERRED FROM 27 VERTICALLY EVOLVING MARKER GENES

The phylogeny of the primary domains of life inferred from the 27 most vertically evolving genes as inferred based on our ranking of markers and using the best-fitting LG + C60 + G4 + F model (Fig. 4) is consistent with recent single-domain trees inferred for Archaea and Bacteria independently (*40*, *44*, *50*), although the deep relationships within Bacteria are poorly resolved, with the exception of the monophyly of Gracilicutes (Fig. 4). Our results are also in good agreement with a recent estimate of the universal tree based on a different marker gene selection approach (*51*). In that study, marker genes were selected based on Tree Certainty, a metric that quantifies phylogenetic signal based on the extent to which markers distinguish between different resolutions of conflicting relationships (*52*).

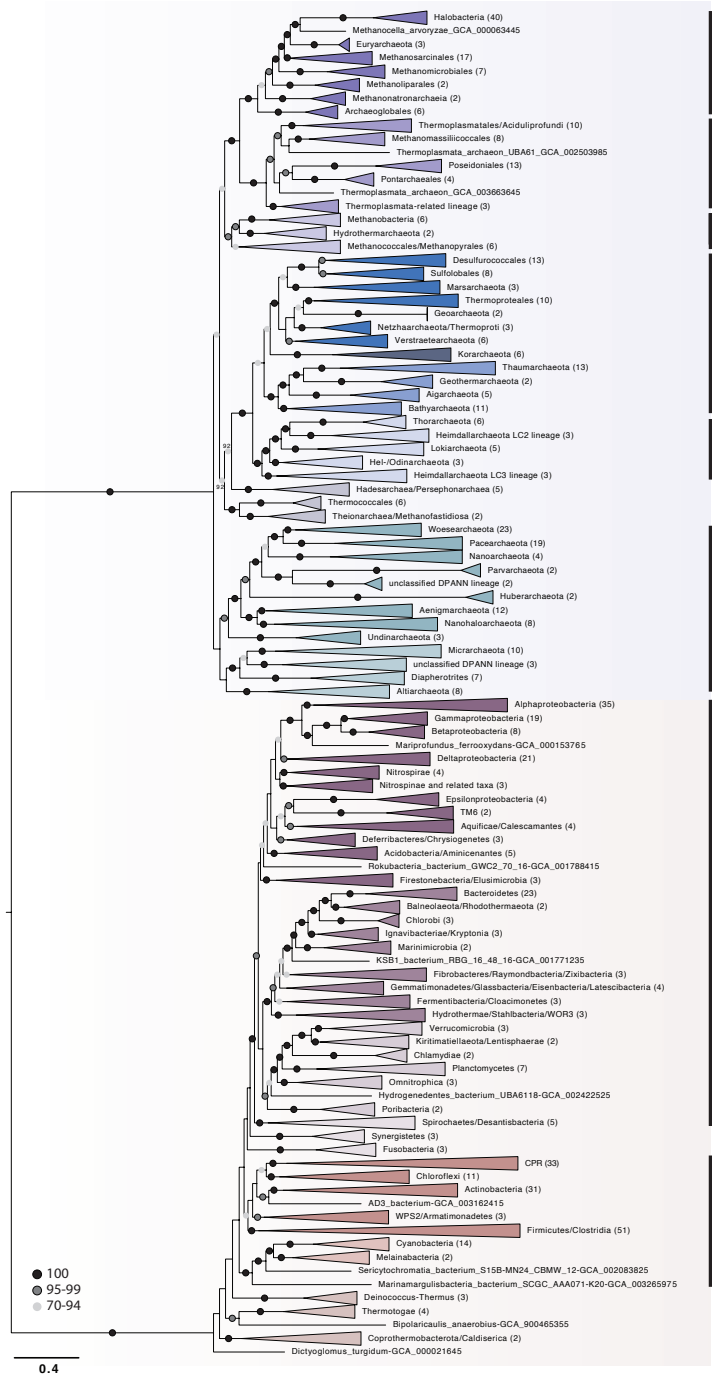


Fig. 4 | A phylogeny of Archaea and Bacteria inferred from a concatenation of 27 marker genes. Consistent with some recent studies (*8*, *44*, *50*, *53*), we recovered the DPANN, TACK, and Asgard Archaea as monophyletic groups. Although the deep branches within Bacteria are poorly resolved, we recovered

a sister group relationship between candidate phyla radiation (CPR) and Chloroflexota, consistent with recent reports (*40*, *54*). The tree was inferred using the best-fitting LG + C60 + G4 + F model in IQ-TREE 2 (*43*). Branch lengths are proportional to the expected number of substitutions per site. Support values are ultrafast (UFBoot2) bootstraps (*55*). Numbers in parenthesis refer to the number of taxa within each collapsed clade. Please note that the collapsed taxa in the Archaea and Bacteria roughly correspond to order- and phylum-level lineages, respectively.

Fig. supplement 1. Raw count and percentage distribution of the Genome Taxonomy Database (GTDB)-defined classes of 350 archaea and 350 bacteria used in the 27 marker gene set analysis.
Fig. supplement 2. Raw count and percentage distribution of the Genome Taxonomy Database (GTDB)-defined phyla of 350 archaea and 350 bacteria used in the 27 marker gene set analysis.
Fig. supplement 3. A phylogeny of Archaea and Bacteria inferred from a concatenation of 25 marker genes.

In particular, our analysis placed the candidate phyla radiation (CPR) (*56*) as a sister lineage to Chloroflexi (Chloroflexota) rather than as a deep-branching bacterial superphylum. While this contrasts with initial trees suggesting that CPR may represent an early diverging sister lineage of all other Bacteria (*4*, *38*, *56*), our finding is consistent with recent analyses that have instead recovered CPR within the Terrabacteria (*40*, *51*, *54*). Together, these analyses suggest that the deep-branching position of CPR in some trees may be a result of long branch attraction, a possibility that has been raised previously (*4*, *57*).

The deep branches of the archaeal subtree are generally well-resolved and recover DPANN (51% bootstrap support), Asgards (100% bootstrap support), and TACK Archaea (75% bootstrap support) as monophyletic clades in agreement with a range of previous studies (*8*, *44*, *50*, *53*). We also find support for the placement of Methanonatronarchaeia (*58*) distant to Halobacteria as one of the earliest branches of the Methanotecta (Fig. 4, Fig. 4—figure supplement 3) in agreement with recent analyses, suggesting that their initial placement with Halobacteria (*58*) may be an artifact of compositional attraction (*44*, *59–61*).

We obtained moderate (92%) bootstrap support for the branching of some Euryarchaeota with the TACK + Asgard clade: the Hadesarchaea + Persephonarchaea were resolved as the sister group to TACK + Asgards with moderate (92%) support, with this entire lineage branching sister to a strongly supported (100%) clade comprising Theionarchaea, Methanofastidiosa, and Thermococcales. However, the position of these lineages was sensitive to the marker gene set used. As part of a robustness test, we also inferred an additional tree from a 25-gene subset, excluding two genes that have complex evolutionary histories in Archaea (*62*); Fig. 4—figure supplement 3). In this analysis, these Archaea instead branched with Methanomada with high support (98%), highlighting the difficulty of placing these lineages in the archaeal tree. Euryarchaeotal paraphyly has been previously reported (*8*, *50*, *63*, *64*), though the extent of the observed paraphyly and the lineages involved has varied among analyses.

A basal placement of DPANN within Archaea is sometimes viewed with suspicion (*65*) because DPANN genomes are reduced and appear to be fast-evolving, properties that may cause LBA artifacts (*66*) when analyses include Bacteria. However, in contrast to CPR, with which DPANN share certain ecological and genomic similarities (e.g., host dependency, small genomes,

limited metabolic potential), the early divergence of DPANN from the archaeal branch has received support from a number of recent studies (*44*, *50*, *64*, *67–70*), though the inclusion of certain lineages within this radiation remains controversial (*60*, *65*) and the placement of the root is uncertain (*44*, *64*). While more in-depth analyses will be needed to further illuminate the evolutionary history of DPANN and establish whether the group as a whole is monophyletic, our work is in agreement with current literature and a recently established phylogeny-informed archaeal taxonomy (*69*).

A broader observation from our analysis is that the phylogenetic diversity of the archaeal and bacterial domains, measured as substitutions per site in this consensus set of vertically evolving marker genes, appears to be similar (Fig. 3A; the mean root-to-tip distance for archaea: 2.38, for bacteria: 2.41; the range of root-to-tip distances for archaea: 1.79–3.01, for bacteria: 1.70–3.17). Considering only the slowest-evolving category of sites, branch lengths within Archaea are actually longer than within Bacteria (Fig. 3C). This result differs from some published trees (*4*, *24*) in which the phylogenetic diversity of Bacteria has appeared to be significantly greater than that of Archaea. By contrast to those earlier studies, we analyzed a set of 350 genomes from each domain, an approach that may tend to reduce the differences between them. While we had to significantly downsample the sequenced diversity of Bacteria, our sampling nonetheless included representatives from all known major lineages of both domains (Fig. 4—figure supplements 1 and 2, see Fig. 1—figure supplements 14–16 for a comparison with the expanded marker set), and so might be expected to recover a difference in diversity, if present. Our analyses and a number of previous studies (*4*, *6*, *24*, *71*) indicate that the choice of marker genes has a profound impact on the apparent phylogenetic diversity of certain prokaryotic groups; for instance, in the proportion of bacterial diversity composed of CPR (*4*, *72*). Our results demonstrate that slow- and fast-evolving sites from the same set of marker genes support different tree shapes and branch lengths; it therefore seems possible that between-dataset differences are due, at least in part, to evolutionary rate variation within and between marker genes.

DIFFICULTIES IN ESTIMATING THE AGE OF THE LAST UNIVERSAL COMMON ANCESTOR

While a consensus may be emerging on the topology of the universal tree, estimates of the ages of the deepest branches, and their lengths in geological time, remain highly uncertain. The fossil record of early life is incomplete and difficult to interpret (*73*), and in this context molecular clock methods provide a means of combining the abundant genetic data available for modern organisms with the limited fossil record to improve our understanding of early evolution (*20*). The 381-gene dataset was suggested to be (*24*) useful for inferring deep divergence times because age estimates of LUCA (last universal common ancestor) from this dataset using a strict molecular clock were in agreement with the geological record: a root (LUCA) age of 3.6–4.2 Ga was inferred from the entire 381-gene dataset, consistent with the earliest fossil evidence for life (*19*, *20*). By contrast, analysis of ribosomal markers alone (*24*) supported a

root age of ~7 Ga, which might be considered implausible because it is older than the age of the Earth and Solar System (with the moon-forming impact occurring ~4.52 Ga; ref. *(74)*; ref. *(75)*).

The published molecular clock analyses (*24*) made use of concatenation-based branch lengths in which topological disagreement among sites is not modeled and are likely to be affected by the impact of nonvertical marker genes and substitutional saturation on branch length estimation discussed above. Consistent with this hypothesis, divergence time inference using the same method on the 5% most-vertical subset of the expanded marker set (as determined by ΔLL; this set of 20 genes includes only one ribosomal protein, see Supplementary file 5a) resulted in age estimates for LUCA that exceed the age of the Earth, 5.6–6.15 Ga (Fig. 5), approaching the age inferred from the ribosomal genes (7.46–8.03 Ga). These results (Fig. 5) suggest that the apparent agreement between the fossil record and divergence times estimated from the expanded gene set may be due, at least in part, to the shortening of the AB branch due to phylogenetic incongruence among marker genes.

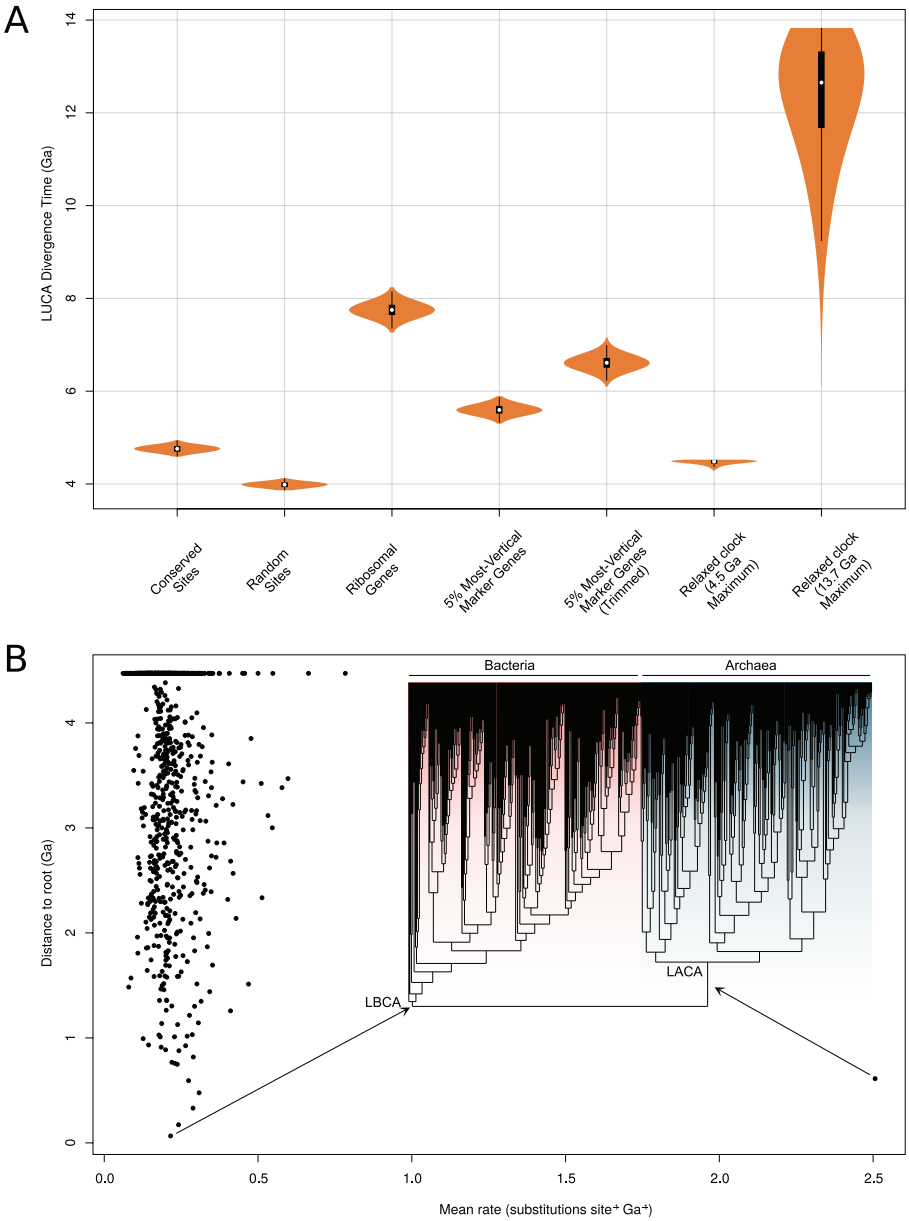


Fig. 5 | Molecular clock estimates of the LUCA and LACA age are uncertain due to a lack of deep calibrations and maximum ages for microbial clades. (**A**) Posterior node age estimates from Bayesian molecular clock analyses of (1) conserved sites as estimated previously (*24*); (2) random sites (*24*); (3) ribosomal genes (*24*); (4) the top 5% of marker gene families according to their ΔLL score (including only one ribosomal protein); and (5) the same top 5% of marker genes trimmed using BMGE (*76*) to remove poorly aligned sites. In each case, a strict molecular clock was applied, with the age of the Cyanobacteria-Melainabacteria split constrained between 2.5 and 2.6 Ga. In (6) all annotations for the top 20 genes and

(7) an expanded set of fossil calibrations was implemented with a relaxed (lognormal) molecular clock. In (6), a soft maximum age of 4.520 Ga was applied, representing the age of the moon-forming impact (*77*). In (7), a soft maximum age corresponding to the estimated age of the universe (*78*) was applied. (**B**) Inferred rates of molecular evolution along the phylogeny in a relaxed clock analysis where the maximum age was set to 4.520 Ga. We plotted per-branch rates (site^{-1} Ga^{-1}) against distance to the root. The rate of evolution along the archaea stem lineage was a clear outlier (mean = 2.51, 95% HPD = 1.6–3.5subs. site^{-1} Ga^{-1}). The phylogeny used was that depicted in Fig. 4.

In the original analyses, the age of LUCA was estimated using a strict clock with a single calibration constraining the split between Cyanobacteria and Melainabacteria derived from estimates of the Great Oxidation Event and a secondary estimate of the age of cyanobacteria derived from an independent analysis (*79*). The combination of a strict clock and only two calibrations is not sufficient to capture the variation in evolutionary rate over deep timescales (*80*). To investigate whether additional calibrations might help to improve age estimates for deep nodes in the universal tree, we performed analyses on our new 27 marker gene dataset using two different relaxed clock models (with branchwise independent and autocorrelated rates) and seven additional calibrations (Supplementary file 5b). Unfortunately, all of these were minimum age calibrations with the exception of the root (for which the moon-forming impact 4.52 Ga (*77*)) provides a reasonable maximum) due to the difficulty of establishing uncontroversial maximum ages for microbial clades. Maximum age constraints are essential to inform faster rates of evolution because, in combination with more abundant minimum age constraints, they imply that a given number of substitutions must have accumulated in at most a certain interval of time. In the absence of other maximum age constraints, the only lower bound on the rate of molecular evolution is provided by the maximum age constraint on the root (LUCA).

These new analyses indicated that even with additional minimum age calibrations the age of LUCA inferred from the 27-gene dataset was unrealistically old, falling close to the maximum age constraint in all analyses even when the maximum was set to the age of the known universe (13.7 Ga; ref. (*78*); Fig. 5). Inspection of the inferred rates of molecular evolution across the tree (Fig. 5B) provides some insight into these results: the mean rate is low (mean = 0.21, 95% credibility interval = 0.19–0.22 subs. site^{-1} Ga^{-1}), so that long branches (such as the AB stem), in the absence of other information, are interpreted as evidence of a long period of geological time. These low rates likely result both from the limited number of calibrations and, in particular, the lack of maximum age constraints.

An interesting outlier among inferred rates is the LUCA to LACA (last archaeal common ancestor) branch, which has a rate 10-fold greater than the average (mean = 2.51, 95% HPD (highest posterior density) = 1.6–3.5 subs. site^{-1} Ga^{-1}). The reason is that calibrations within Bacteria imply that LBCA (last bacterial common ancestor) cannot be younger than 3.225 Ga (Manzimnyama Banded Ironstone Formation provides evidence of cyanobacterial oxygenation; ref. (*81*), Supplementary file 5b); as a result, with a 4.52 Ga maximum, the LUCA to LBCA branch cannot be longer than 1.295 Ga. By contrast, the early branches of the archaeal tree are poorly constrained by fossil evidence. Analysis without the 3.225 Ga constraint resulted

in overlapping age estimates for LBCA (4.47–3.53 Ga) and LACA (4.37–3.44 Ga). Finally, analysis of the archaeal and bacterial subtrees independently (i.e., without the AB branch, rooted on LACA and LBCA, respectively) resulted in LBCA and LACA ages that abut the maximum root age (LBCA: 4.52–4.38 Ga; LACA: 4.52–4.14 Ga). This analysis demonstrates that, under these analysis conditions, the inferred age of the root (whether corresponding to LUCA, LACA, or LBCA) is strongly influenced by the prior assumptions about the maximum age of the root.

In sum, the agreement between fossils and age estimates from the expanded gene set appears to result from the impact of phylogenetic incongruence on branch length estimates. Under more flexible modeling assumptions, the limitations of current clock methods for estimating the age of LUCA become manifest: the sequence data only contain limited information about the age of the root, with posterior estimates driven by the prior assumptions about the maximum age of the root. This analysis implies several possible ways to improve age estimates of deep branches in future analyses. More calibrations, particularly maximum age constraints and calibrations within Archaea, are essential to refine the current estimates. Given the difficulties in establishing maximum ages for archaeal and bacterial clades, constraints from other sources such as donor-recipient age constraints inferred from HGTs (*82*–*85*), or clock models that capture biological opinion about rate shifts in early evolution, may be particularly valuable.

CONCLUSION

Our analysis of a range of published marker gene datasets (*6*, *10*, *24*, *27*) indicates that the choice of markers and the fit of the substitution model are both important for inference of deep phylogeny from concatenations, in agreement with an existing body of literature (reviewed in refs. (*34*, *86*, *87*)). We established a set of 27 vertically evolving marker gene families and found no evidence that ribosomal genes overestimate stem length; since they appear to be transferred less frequently than other genes, our analysis affirms that ribosomal proteins are useful markers for deep phylogeny. In general, high-verticality markers, regardless of functional category, supported a longer AB branch length. Furthermore, our phylogeny was consistent with recent work on early prokaryotic evolution, resolving the major clades within Archaea and nesting the CPR within Terrabacteria. Notably, our analyses suggested that both the true AB branch length (Fig. 6A) and the phylogenetic diversity of Archaea may be underestimated by even the best current models, a finding that is consistent with a root for the tree of life between the two prokaryotic domains. Taken together with fossil evidence for crown-group Bacteria ~3.2 Ga (*20*), the long AB branch length inferred from vertically evolving genes is consistent with the hypothesis that rates of molecular evolution may have been higher early in life's history than more recently (*88*), although the inferred rates are uncertain and contingent on limited fossil evidence and the assumptions of the molecular clock model.

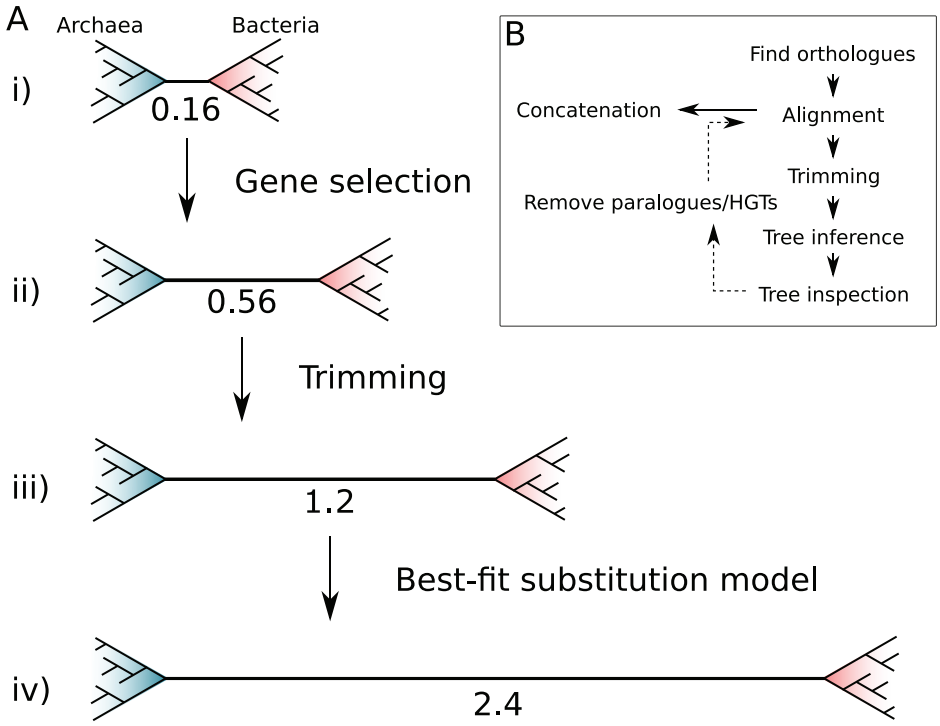


Fig. 6 | **The impact of marker gene choice, phylogenetic congruence, alignment trimming, and substitution model fit on estimates of the Archaea-Bacteria (AB) branch length.** (**A**) Analysis using a site-homogeneous model (LG + G4 + F) on the complete 381-gene expanded set (i) results in an AB branch substantially shorter than previous estimates. Removing the genes most seriously affected by inter-domain gene transfer (ii), trimming poorly aligned sites (iii) using BMGE (*76*) in the original alignments (see below), and using the best-fitting site-heterogeneous model (iv) (LG + C60 + G4 + F) substantially increase the estimated AB length, such that it is comparable with published estimates from the 'core' set: 3.3 (*10*) and the consensus set of 27 markers identified in the present study: 2.5. Branch lengths measured in expected number of substitutions/site. (**B**) Workflow for iterative manual curation of marker gene families for concatenation analysis. After inference and inspection of initial ortholog trees, several rounds of manual inspection and removal of HGTs and distant paralogs were carried out. These sequences were removed from the initial set of orthologs before alignment and trimming. For a detailed discussion of some of these issues, and practical guidelines on phylogenomic analysis of multi-gene datasets, see ref. (*87*) for a useful review.

Phylogenies inferred from 'core' genes involved in translation and other conserved cellular processes have provided one of the few available windows into the earliest period of archaeal and bacterial evolution. However, core genes comprise only a small proportion of prokaryotic genomes and have sometimes been viewed as outliers (*24*) in the sense that they are unusually vertical among prokaryotic gene families. This means that they are among the few prokaryotic gene families amenable to concatenation methods, which are useful for pooling signal from individual weakly resolved gene trees but which make the assumption that all sites evolve on the same underlying tree. If other gene families are included in concatenations, the results can be difficult to predict because differences in topology across sites are not modeled.

Our analyses of the 381-gene expanded set suggest that this incongruence can lead to underestimation of the evolutionary distance between Archaea and Bacteria, in the sense of the branch length separating the archaeal and bacterial domains. We note that alternative conceptions of evolutionary distance are possible; for example, in a phenetic sense of overall genome similarity, extensive HGT will increase the evolutionary proximity (*24*) of the domains so that Archaea and Bacteria may become intermixed at the single-gene level. While such data can encode an important evolutionary signal, it is not amenable to concatenation analysis. At the same time, it is clearly unsatisfactory to base our view of early evolution on a relatively small set of genes that appear to experience selective pressures rather distinct from the forces at play more broadly in prokaryotic genome evolution. These limitations are particularly unfortunate given the wealth of genome data now available to test hypotheses about early evolution. Exploring the evolutionary signal in more of the genome than hitherto is clearly a worthwhile endeavor. New methods, including more realistic models of gene duplication, transfer and loss (*89*, *90*), and extensions to supertree methods to model paralogy (*91*) and gene transfer, promise to enable genome-wide inference of prokaryotic history and evolutionary processes using methods that can account for the varying evolutionary histories of individual gene families.

MATERIALS AND METHODS

DATA

We downloaded the individual alignments from Zhu et al., 2019 (*24*) (https://github.com/biocore/wol/tree/master/data/, Zhu, 2022 (*92*)), along with the genome metadata and the individual Newick files. We checked each published tree for domain monophyly and also performed AU (*39*) tests to assess support for domain monophyly on the underlying sequence alignments using IQ-TREE 2.0.6 (*43*). The phylogenetic analyses were carried out using the 'reduced' subset of 1000 taxa outlined by the authors (*24*) for computational tractability. These markers were trimmed according to the protocol in the original paper (*24*), that is, sites with >90% gaps were removed, followed by removal of sequences with >66% gaps.

We also downloaded the Williams et al., 2020 (*10*) ('core'), Petitjean et al., 2014 (*6*) ('non-ribosomal'), and Coleman et al., 2021 (*40*) ('bacterial') datasets from their original publications.

ANNOTATIONS

Proteins used for phylogenetic analyses by Zhu et al., 2019 (*24*) were annotated to investigate the selection of sequences comprising each of the marker gene families. To this end, we downloaded the protein sequences provided by the authors from the following repository: https://github.com/biocore/wol/tree/master/data/alignments/genes. To obtain reliable annotations, we analyzed all sequences per gene family using several published databases, including the arCOGs (version from 2014) (*93*), KOs from the KEGG Automatic Annotation Server (KAAS; downloaded April 2019) (*94*), the PFAM database (release 31.0)

(*95*), the TIGRFAM database (release 15.0) (*96*), the Carbohydrate-Active enZymes (CAZy) database (downloaded from dbCAN2 in September 2019) (*97*), the MEROPs database (release 12.0) (*98*, *99*), the hydrogenase database (HydDB; downloaded in November 2018) (*100*), the NCBI_ non-redundant (nr) database (downloaded in November 2018), and the NCBI COGs database (version from 2020). Additionally, all proteins were scanned for protein domains using InterProScan (v5.29–68.0; settings: --iprlookup --goterms) (*101*).

Individual database searches were conducted as follows: arCOGs were assigned using PSI-BLAST v2.7.1+ (settings: -evalue 1e-4 -show_gis -outfmt 6 -max_target_seqs 1000 -dbsize 100000000 -comp_based_stats F -seg no) (*102*). KOs (settings: -E 1e-5), PFAMs (settings: -E 1e-10), TIGRFAMs (settings: -E 1e-20), and CAZymes (settings: -E 1e-20) were identified in all archaeal genomes using hmmsearch v3.1b2 (*103*). The MEROPs and HydDB databases were searched using BLASTp v2.7.1 (settings: -outfmt 6, -evalue 1e-20). Protein sequences were searched against the NCBI_nr database using DIAMOND v0.9.22.123 (settings: –more-sensitive –e-value 1e-5 –seq 100 –no-self-hits –taxonmap prot.accession2taxid.gz) (*104*). For all database searches, the best hit for each protein was selected based on the highest e-value and bitscore and all results are summarized in Supplementary file 1 and full results are given in the Data Supplement Expanded_Bacterial_Core_Nonribosomal_analyses/Annotation_Tables/0_Annotation_tables_full/All_Zhu_marker_annotations_16-12-2020.tsv.zip. For InterProScan, we report multiple hits corresponding to the individual domains of a protein using a custom script (parse_IPRdomains_vs2_GO_2.py).

Assigned sequence annotations were summarized, and all distinct KOs and PFAMs were collected and counted for each marker gene. KOs and PFAMs with their corresponding descriptions were mapped to the marker gene file downloaded from the repository here: (https://github.com/biocore/wol/blob/master/data/markers/metadata.xlsx) and used in summarization of the 381 marker gene protein trees (Supplementary file 1).

For manual inspection of single marker gene trees, KO and PFAM annotations were mapped to the tips of the published marker protein trees, downloaded from the repository here: (https://github.com/biocore/wol/tree/master/data/trees/genes). Briefly, the Genome ID, PFAM, PFAM description, KO, KO description, and NCBI Taxonomy string were collected from each marker gene annotation table and were used to generate mapping files unique to each marker gene phylogeny, which links the Genome ID to the annotation information (GenomeID|Domain|Pfam|Pfam Description|KO|KO Description). An in-house Perl script replace_tree_names.pl (available here: https://github.com/ndombrowski/Phylogeny_tutorial/tree/main/Input_files/5_required_scripts; Dombrowski, 2022 (*105*) copy archived at: https://archive.softwareheritage.org/browse/directory/519c8e6f6a054ed312f0c1a311e6fda461ec189f/?origin_url=https://github.com/ndombrowski/Phylogeny_tutorial&revision=59ce418ec42160a15e82610120220b611b6e96db&snapshot=48497545c8d19a8288c1e02a21ea284bb4ae1671) was used to append the summarized protein annotations to the corresponding tips in each marker gene tree. Annotated marker gene phylogenies were manually inspected

using the following criteria, including (1) retention of reciprocal domain monophyly (Archaea and Bacteria) and (2) for the presence or absence of potential paralogous families. Paralogous groups and misannotated families present in the gene trees were highlighted and violations of search criteria were recorded in Supplementary file 1.

PHYLOGENETIC ANALYSES

COG assignment for the core, non-ribosomal, and bacterial marker genes

First, all gene sequences in the three published marker sets (core, non-ribosomal, and bacterial) were annotated using the NCBI COGs database (version from 2020). Sequences were assigned a COG family using hmmsearch v3.3.2 (*103*) (settings: -E 1e-5) and the best hit for each protein sequence was selected based on the highest e-value and bit score. To assign the appropriate COG family for each marker gene, we quantified the percentage distribution of all unique COGs per gene and selected the family representing the majority of sequences in each marker gene.

Accounting for overlap, this resulted in 95 unique COG families from the original 119 total marker genes across all three published datasets (Supplementary file 2). Orthologs corresponding to these 95 COG families were identified in the 700 genomes (350 Archaea, 350 Bacteria, Supplementary file 3) using hmmsearch v3.3.2 (settings: -E 1e-5). The reported BinID and protein accession were used to extract the sequences from the 700 genomes, which were used for subsequent phylogenetic analyses.

Marker gene inspection and analysis

We aligned these 95 marker gene sequence sets using MAFFT-L-INS-i 7.475 (*106*) and removed poorly aligned positions with BMGE 1.12 (*76*). We inferred initial ML trees (LG + G4 + F) for all 95 markers and mapped the KO and PFAM domains and descriptions, inferred from annotation of the 700 genomes, to the corresponding tips (see above). Manual inspection took into consideration monophyly of Archaea and Bacteria and the presence of paralogs, and other signs of contamination (HGT, LBA). Accordingly, single-gene trees that failed to meet reciprocal domain monophyly were excluded, and any instances of HGT, paralogous sequences, and LBA artifacts were manually removed from the remaining trees, resulting in 54 markers across the three published datasets that were subject to subsequent phylogenetic analysis (LG + C20 + G4 + F) and further refinement (see below).

Ranking markers based on split score

We applied an automated marker gene ranking procedure devised previously (the split score, Dombrowski et al., 2020 (*44*)) to rank each of the 54 markers that satisfied reciprocal monophyly based on the extent to which they recovered established phylum-, class-, or order-level relationships within the archaeal and bacterial domains (Supplementary file 4).

The script quantifies the number of splits, or occurrences where a taxon fails to cluster within its expected taxonomic lineage, across all gene phylogenies. Briefly, we assessed

monophyletic clustering using phylum-, class-, and order-level clades within Archaea (Cluster1) in combination with Cluster0 (phylum) or Cluster3 (i.e., on class-level if defined and otherwise on phylum-level; Supplementary file 4) for Bacteria. We then ranked the marker genes using the following split score criteria: the number of splits per taxon and the splits normalized to the species count. The percentage of split phylogenetic groups was used to determine the highest ranking (top 50%) markers.

Concatenation

Based on the split score ranking of the 54 marker genes (above), the top 50% (27 markers, Supplementary file 4) marker genes were manually inspected using criteria as defined above, and contaminating sequences were manually removed from the individual sequence files. Following inspection, marker protein sequences were aligned using MAFFT-L-INS-i 7.475 (*107*) and trimmed using BMGE (version 1.12, under default settings) (*76*). We concatenated the 27 markers into a supermatrix, which was used to infer an ML tree (Fig. 4, under LG + C60 + G4 + F), evolutionary rates (see below), and rate category supermatrices, as well as to perform model performance tests (see below). We also concatenated the non-ribosomal and ribosomal markers from the 27 and 54 marker sets into four more supermatrices and inferred ML trees under (LG + C60 + G4 + F) (Table 1). Two additional supermatrices were constructed from the 54 markers, one before manual removal of apparent HGTs and one after the removal, with both sets of markers aligned and trimmed in the same way as the other datasets (see above). We also inferred an ML tree under LG + C60+ G4 + F from a supermatrix consisting of a concatenation of 25 marker genes after removing COG0480 and COG5257.

CONSTRAINT ANALYSIS

We performed an ML free topology search using IQ-TREE 2.0.6 (*43*) under the LG + G4 + F model, with 1000 ultrafast bootstrap replicates (*55*) on each of the markers from the expanded, bacterial, core, and non-ribosomal sets. We also performed a constrained analysis with the same model in order to find the ML tree in which Archaea and Bacteria were reciprocally monophyletic. For the expanded set, we plotted branch lengths from the maximum likelihood trees constrained to recover domain monophyly; for the other datasets, we plotted branch lengths from the maximum likelihood trees. We then compared both trees using the AU (*39*) test in IQ-TREE 2.0.6 (*43*) with 10,000 RELL (*39*) bootstrap replicates. To evaluate the relationship between marker gene verticality and AB branch length, we calculated the difference in log-likelihood between the constrained and unconstrained trees in order to rank the genes from the expanded marker set. We then concatenated the top 20 markers (with the lowest difference in log-likelihood between the constrained and unconstrained trees) and iteratively added five markers with the next smallest difference in log-likelihood to the concatenate; this was repeated until we had concatenates up to 100 markers (with the lowest difference in log-likelihood) we inferred trees under LG + C10 + G4 + F in IQ-TREE 2.0.6, with 1000 ultrafast bootstrap replicates and calculated AB length.

SITE AND GENE EVOLUTIONARY RATES

We inferred rates using the --rate option in IQ-TREE 2.0.6 (*43*) for both the 381 marker concatenation from Zhu et al., 2019 (*24*) and the top 5% of marker genes based on the results of difference in log-likelihood between the constrained tree and free-tree search in the constraint analysis (above). We also used this method to explore the differences in rates for the 27 marker set. We built concatenates for sites in the slowest and fastest rate categories, and inferred branch lengths from each of these concatenates using the tree inferred from the corresponding dataset as a fixed topology.

SUBSTITUTION MODEL FIT

Model fit tests were undertaken using the top 5% concatenate described above, with the alignment being trimmed with BMGE 1.12 (*76*) with default settings (BLOSUM62, entropy 0.5) for all of the analyses except the 'untrimmed' LG + G4 + F run; other models on the trimmed alignment were LG + G4 + F, LG + R4 + F and LG + C10,20,30,40,50,60 + G4 + F, with 1000 ultrafast (*55*) bootstrap replicates. Model fitting was done using ModelFinder (*108*) in IQ-TREE 2.0.6 (*43*). For the 27 marker concatenation, we performed a model finder analysis (-m MFP) including additional complex models of evolution (i.e., LG + C60 + G4 + F, LG + C50 + G4 + F, LG + C40 + G4 + F, LG + C30 + G4 + F, LG + C20 + G4 + F, LG + C10 + G4 + F, LG + G4 + F, LG + R4+ F) to the default, to find the best-fitting model for the analysis. This revealed that, according to AIC (Akaike information criterion), BIC (Bayesian information criterion), and cAIC (corrected Akaike information criterion), LG + C60 + G4 + F was the best-fitting model. For comparison, we also performed analyses using the following models: LG + G4 + F, LG + C20 + G4 + F, LG + C40 + G4 + F (Table 1).

MOLECULAR CLOCK ANALYSES

Molecular clock analyses were devised to test the effect of genetic distance on the inferred age of LUCA. Following the approach of Zhu et al., 2019 (*24*), we subsampled the alignment to 100 species. Five alternative alignments were analyzed, representing conserved sites across the entire alignment, randomly selected sites across the entire alignment, only ribosomal marker genes, the top 5% of marker genes according to ΔLL, and the top 5% of marker genes further trimmed under default settings in BMGE 1.12 (*76*). Divergence time analyses were performed in MCMCTree (*109*) under a strict clock model. We used the normal approximation approach, with branch lengths estimated in codeml under the LG + G4 model. In each case, a fixed tree topology was used alongside a single calibration on the Cyanobacteria-Melainabacteria split. The calibration was modeled as a uniform prior distribution between 2.5 and 2.6 Ga, with a 2.5% probability that either bound could be exceeded. For each alignment, four independent MCMC chains were run for 2,000,000 generations to achieve convergence.

We repeated clock analyses under a relaxed (independent rates drawn from a lognormal distribution) clock model with an expanded sampling of fossil calibration (Supplementary file 5b). We repeated the analyses with two approaches to define the maximum age calibration. The first used the moon-forming impact (4.52 Ga) under the provision that no forms of life

are likely to have survived this event. The second relaxed this assumption, instead using the estimated age of the universe (13.7 Ga) as a maximum. Analyses were performed as above.

SPLIT SCORE ANALYSIS FOR EXPANDED SET MARKERS

We used the previously described split score ranking procedure to quantify the number of taxonomic splits in the 381 marker gene phylogenies generated using the 1000-taxa subsample defined by Zhu et al., 2019 (*24*). Taxonomic clusters were assigned using the Genome Taxonomy Database (GTDB) taxonomic ranks downloaded from the repository (https://github.com/biocore/wol/tree/master/data/taxonomy/gtdb). Lineage-level monophyly was defined at the class level for all archaea (Arc1) and the phylum level for all bacteria (Bac0) (Supplementary file 1).

Of the original 10,575 genomes, 843 lacked corresponding GTDB assignments. For complete taxonomic coverage of the dataset, we used the GTDB Toolkit (GTDB-Tk) v0.3.2 (*110*) to classify these genomes based on GTDB release 202. One of the 843 unclassified taxa (gid: G000715975) failed the GTDB-Tk quality control check, resulting in no assignment; therefore, we manually assigned this taxon to the Actinobacteriota based on the corresponding affiliation to the Actinobacteria in the NCBI taxonomic ranks provided in the genomic metadata downloaded from the repository (https://github.com/biocore/wol/tree/master/data/genomes). Additionally, two archaeal taxa within the Poseidoniia_A (gids: G001629155, G001629165) were manually assigned to the archaeal class MGII (Supplementary file 1).

PLOTTING

Split score statistical analyses were performed using R 3.6.3 (R Core Team, 2020). All other statistical analyses were performed using R 4.0.4 (*111*), and data were plotted with ggplot2 (*112*).

ACKNOWLEDGEMENTS

This work was supported by the Gordon and Betty Moore Foundation through grant GBMF9741 to TAW, AS, and GJSz. ERRM was supported by a Royal Society Enhancement Award (RGF\EA\180199) to TAW. CP was supported by NERC grant NE/P00251X/1 to TAW. TAW was supported by a Royal Society University Research Fellowship (URF\R\201024). GJSz received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program under Grant Agreement 714774 and Grant GINOP-2.3.2.-15-2016-00057. AS received funding from the Swedish Research Council (VR starting grant 2016-03559), the NWO-I foundation of the Netherlands Organisation for Scientific Research (WISE fellowship), and the European Research Council (ERC Starting grant 947317, ASymbEL). ND was supported through the WISE fellowship, ERC StG 947317 and GBMF9741 to AS.

AUTHOR CONTRIBUTIONS

Edmund RR Moody, Tara A Mahendrarajah, Nina Dombrowski, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review and editing; James W Clark, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review and editing; Celine Petitjean, Data curation, Formal analysis, Writing – review and editing; Pierre Offre, Conceptualization, Data curation, Formal analysis, Writing – review and editing; Gergely J Szöllősi, Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review and editing; Anja Spang, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Supervision, Writing – original draft, Writing – review and editing; Tom A Williams, Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Supervision, Writing – original draft, Writing – review and editing

ADDITIONAL FILES

Supplementary files

- Supplementary file 1. Marker metadata, KO and Pfam annotations and descriptions, and manual inspection notes for reciprocal monophyly and presence of paralogs for 381 marker genes used in ***Zhu et al., 2019*** *(24)* (Materials and methods).
- Supplementary file 2. Marker metadata, KO and Pfam annotations and descriptions, and manual inspection notes for 95 markers in the core, bacterial, and non-ribosomal marker gene sets (Materials and methods).
- Supplementary file 3. NCBI taxonomic information for 350 archaeal and 350 bacterial genomes sampled in the new analyses.
- Supplementary file 4. Clade definitions for quantifying taxonomic splits and split score statistical summaries for ranking of the core, bacterial, non-ribosomal marker genes, and 381 marker genes (Materials and methods).
- Supplementary file 5. Annotations of the top 20 genes from the expanded set, and a list of fossil calibrations. (a) Functional annotations for the top 20 genes used and in ***Fig. 6*** and referred to in ***Fig. 6A***. (b) A list of fossil calibrations employed in relaxed molecular clock analyses. All calibrations were modeled as uniform distributions between a hard minimum and a soft maximum. The probability that the maximum could be exceeded was modeled as a 2.5% probability tail.
- Transparent reporting form

All figure supplements for Figs. 1-4 and Supplementary files can be accessed here: https://elifesciences.org/articles/66695/figures#supp

Data availability

All of the data, including sequence alignments, trees, annotation files, and scripts associated with this manuscript have been deposited in the FigShare repository at https://doi.org/10.6084/m9.figshare. 13395470.

REFERENCES

1. F. D. Ciccarelli, T. Doerks, C. von Mering, C. J. Creevey, B. Snel, P. Bork, Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287 (2006).

2. G. P. Fournier, J. P. Gogarten, Rooting the ribosomal tree of life. *Mol. Biol. Evol.* **27**, 1792–1801 (2010).

3. J. K. Harris, S. T. Kelley, G. B. Spiegelman, N. R. Pace, The genetic core of the universal ancestor. *Genome Res.* **13**, 407–412 (2003).

4. L. A. Hug, B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D. A. Relman, K. M. Finstad, R. Amundson, B. C. Thomas, J. F. Banfield, A new view of the tree of life. *Nat Microbiol* **1**, 16048 (2016).

5. S. Mukherjee, R. Seshadri, N. J. Varghese, E. A. Eloe-Fadrosh, J. P. Meier-Kolthoff, M. Göker, R. C. Coates, M. Hadjithomas, G. A. Pavlopoulos, D. Paez-Espino, Y. Yoshikuni, A. Visel, W. B. Whitman, G. M. Garrity, J. A. Eisen, P. Hugenholtz, A. Pati, N. N. Ivanova, T. Woyke, H.-P. Klenk, N. C. Kyrpides, 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat. Biotechnol.* **35**, 676–683 (2017).

6. C. Petitjean, P. Deschamps, P. López-García, D. Moreira, Rooting the domain archaea by phylogenomic analysis supports the foundation of the new kingdom Proteoarchaeota. *Genome Biol. Evol.* **7**, 191–204 (2014).

7. H. G. Ramulu, M. Groussin, E. Talla, R. Planel, V. Daubin, C. Brochier-Armanet, Ribosomal proteins: toward a next generation standard for prokaryotic systematics? *Mol. Phylogenet. Evol.* **75**, 103–117 (2014).

8. K. Raymann, C. Brochier-Armanet, S. Gribaldo, The two-domain tree of life is linked to a new root for the Archaea. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 6670–6675 (2015).

9. D. L. Theobald, A formal test of the theory of universal common ancestry. *Nature* **465**, 219–222 (2010).

10. T. A. Williams, C. J. Cox, P. G. Foster, G. J. Szöllősi, T. M. Embley, Phylogenomics provides robust support for a two-domains tree of life. *Nat Ecol Evol* **4**, 138–147 (2020).

11. E. V. Koonin, Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.* **1**, 127–136 (2003).

12. T. Dagan, W. Martin, The tree of one percent. *Genome Biol.* **7**, 118 (2006).

13. C. J. Creevey, T. Doerks, D. A. Fitzpatrick, J. Raes, P. Bork, Universally distributed single-copy genes indicate a constant rate of horizontal transfer. *PLoS One* **6**, e22099 (2011).

14. P. Puigbò, Y. I. Wolf, E. V. Koonin, Search for a "Tree of Life" in the thicket of the phylogenetic forest. *J. Biol.* **8**, 59 (2009).

15. C. J. Cox, P. G. Foster, R. P. Hirt, S. R. Harris, T. M. Embley, The archaebacterial origin of eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 20356–20361 (2008).

16. J. P. Gogarten, H. Kibak, P. Dittrich, L. Taiz, E. J. Bowman, B. J. Bowman, M. F. Manolson, R. J. Poole, T. Date, T. Oshima, J. Konishi, K. Denda, M. Yoshida, Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 6661–6665 (1989).

17. N. Iwabe, K. Kuma, M. Hasegawa, S. Osawa, T. Miyata, Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 9355–9359 (1989).

18. G. Pühler, H. Leffers, F. Gropp, P. Palm, H. P. Klenk, F. Lottspeich, R. A. Garrett, W. Zillig, Archaebacterial DNA-dependent RNA polymerases testify to the evolution of the eukaryotic nuclear genome. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 4569–4573 (1989).

19. K. Sugitani, K. Mimura, M. Takeuchi, K. Lepot, S. Ito, E. J. Javaux, Early evolution of large micro-organisms with cytological complexity revealed by microanalyses of 3.4 Ga organic-walled microfossils. *Geobiology* **13**, 507–521 (2015).

20. H. C. Betts, M. N. Puttick, J. W. Clark, T. A. Williams, P. C. J. Donoghue, D. Pisani, Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nat Ecol Evol* **2**, 1556–1562 (2018).

21. J. Horita, M. E. Berndt, Abiogenic methane formation and isotopic fractionation under hydrothermal conditions. *Science* **285**, 1055–1057 (1999).

22. A. Lepland, G. Arrhenius, D. Cornell, Apatite in early Archean Isua supracrustal rocks, southern West Greenland: its origin, association with graphite and potential as a biomarker. *Precambrian Res.* **118**, 221–241 (2002).

23. M. A. van Zuilen, A. Lepland, G. Arrhenius, Reassessing the evidence for the earliest traces of life. *Nature* **418**, 627–630 (2002).

24. Q. Zhu, U. Mai, W. Pfeiffer, S. Janssen, F. Asnicar, J. G. Sanders, P. Belda-Ferre, G. A. Al-Ghalith, E. Kopylova, D. McDonald, T. Kosciolek, J. B. Yin, S. Huang, N. Salam, J.-Y. Jiao, Z. Wu, Z. Z. Xu, K. Cantrell, Y. Yang, E. Sayyari, M. Rabiee, J. T. Morton, S. Podell, D. Knights, W.-J. Li, C. Huttenhower, N. Segata, L. Smarr, S. Mirarab, R. Knight, Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477 (2019).

25. S. Mirarab, R. Reaz, M. S. Bayzid, T. Zimmermann, M. S. Swenson, T. Warnow, ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541-8 (2014).

26. R. E. Valas, P. E. Bourne, The origin of a derived superkingdom: how a gram-positive bacterium crossed the desert to become an archaeon. *Biol. Direct* **6**, 16 (2011).

27. A. Spang, J. H. Saw, S. L. Jørgensen, K. Zaremba-Niedzwiedzka, J. Martijn, A. E. Lind, R. van Eijk, C. Schleper, L. Guy, T. J. G. Ettema, Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).

28. O. Jeffroy, H. Brinkmann, F. Delsuc, H. Philippe, Phylogenomics: the beginning of incongruence? *Trends Genet.* **22**, 225–231 (2006).

29. P. G. Foster, Modeling compositional heterogeneity. *Syst. Biol.* **53**, 485–495 (2004).

30. N. Lartillot, H. Brinkmann, H. Philippe, Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.* **7 Suppl 1**, S4 (2007).

31. N. Lartillot, H. Philippe, A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).

32. L. S. Quang, O. Gascuel, N. Lartillot, Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **24**, 2317–2323 (2008).

33. H.-C. Wang, K. Li, E. Susko, A. J. Roger, A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC Evol. Biol.* **8**, 331 (2008).

34. T. A. Williams, D. Schrempf, G. J. Szöllősi, C. J. Cox, P. G. Foster, T. M. Embley, Inferring the Deep Past from Molecular Data. *Genome Biol. Evol.* **13** (2021).

35. R. Gouy, D. Baurain, H. Philippe, Rooting the tree of life: the phylogenetic jury is still out. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140329 (2015).

36. N. J. Tourasse, M. Gouy, Accounting for evolutionary rate variation among sequence sites consistently changes universal phylogenies deduced from rRNA and protein-coding genes. *Mol. Phylogenet. Evol.* **13**, 159–168 (1999).

37. N. Segata, D. Börnigen, X. C. Morgan, C. Huttenhower, PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat. Commun.* **4**, 2304 (2013).

38. C. J. Castelle, J. F. Banfield, Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. *Cell* **172**, 1181–1197 (2018).

39. H. Shimodaira, An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508 (2002).

40. G. A. Coleman, A. A. Davín, T. A. Mahendrarajah, L. L. Szánthó, A. Spang, P. Hugenholtz, G. J. Szöllősi, T. A. Williams, A rooted phylogeny resolves early bacterial evolution. *Science* **372** (2021).

41. V. Da Cunha, M. Gaia, D. Gadelle, A. Nasir, P. Forterre, Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS Genet.* **13**, e1006810 (2017).

42. T. A. Williams, P. G. Foster, T. M. W. Nye, C. J. Cox, T. M. Embley, A congruent phylogenomic signal places eukaryotes within the Archaea. *Proc. Biol. Sci.* **279**, 4870–4879 (2012).

43. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, R. Lanfear, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

44. N. Dombrowski, T. A. Williams, J. Sun, B. J. Woodcroft, J.-H. Lee, B. Q. Minh, C. Rinke, A. Spang, Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. *Nat. Commun.* **11**, 3939 (2020).

45. F. D. K. Tria, G. Landan, T. Dagan, Phylogenetic rooting using minimal ancestor deviation. *Nat Ecol Evol* **1**, 193 (2017).

46. M. Y. Galperin, D. M. Kristensen, K. S. Makarova, Y. I. Wolf, E. V. Koonin, Microbial genome analysis: the COG approach. *Brief. Bioinform.* **20**, 1063–1070 (2019).

47. Y. Liu, K. S. Makarova, W.-C. Huang, Y. I. Wolf, A. N. Nikolskaya, X. Zhang, M. Cai, C.-J. Zhang, W. Xu, Z. Luo, L. Cheng, E. V. Koonin, M. Li, Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature* **593**, 553–557 (2021).

48. D. Schrempf, N. Lartillot, G. Szöllősi, Scalable Empirical Mixture Models That Account for Across-Site Compositional Heterogeneity. *Mol. Biol. Evol.* **37**, 3616–3631 (2020).

49. H.-C. Wang, B. Q. Minh, E. Susko, A. J. Roger, Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Syst. Biol.* **67**, 216–235 (2018).

50. T. A. Williams, G. J. Szöllősi, A. Spang, P. G. Foster, S. E. Heaps, B. Boussau, T. J. G. Ettema, T. M. Embley, Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E4602–E4611 (2017).

51. C. A. Martinez-Gutierrez, F. O. Aylward, Phylogenetic Signal, Congruence, and Uncertainty across Bacteria and Archaea. *Mol. Biol. Evol.* **38**, 5514–5527 (2021).

52. L. Salichos, A. Rokas, Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497**, 327–331 (2013).

53. L. Guy, T. J. G. Ettema, The archaeal "TACK" superphylum and the origin of eukaryotes. *Trends Microbiol.* **19**, 580–587 (2011).

54. N. Taib, D. Megrian, J. Witwinowski, P. Adam, D. Poppleton, G. Borrel, C. Beloin, S. Gribaldo, Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. *Nat Ecol Evol* **4**, 1661–1672 (2020).

55. D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

56. C. T. Brown, L. A. Hug, B. C. Thomas, I. Sharon, C. J. Castelle, A. Singh, M. J. Wilkins, K. C. Wrighton, K. H. Williams, J. F. Banfield, Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).

57. R. Méheust, D. Burstein, C. J. Castelle, J. F. Banfield, The distinction of CPR bacteria from other bacteria based on protein family content. *Nat. Commun.* **10**, 4173 (2019).

58. D. Y. Sorokin, K. S. Makarova, B. Abbas, M. Ferrer, P. N. Golyshin, E. A. Galinski, S. Ciordia, M. C. Mena, A. Y. Merkel, Y. I. Wolf, M. C. M. van Loosdrecht, E. V. Koonin, Discovery of extremely halophilic, methyl-reducing euryarchaea provides insights into the evolutionary origin of methanogenesis. *Nat Microbiol* **2**, 17081 (2017).

59. M. Aouad, G. Borrel, C. Brochier-Armanet, S. Gribaldo, Evolutionary placement of Methanonatronarchaeia, *Nature microbiology*. **4** (2019)pp. 558–559.

60. Y. Feng, U. Neri, S. Gosselin, A. S. Louyakis, R. T. Papke, U. Gophna, J. P. Gogarten, The Evolutionary Origins of Extreme Halophilic Archaeal Lineages. *Genome Biol. Evol.* **13** (2021).

61. J. Martijn, M. E. Schön, A. E. Lind, J. Vosseberg, T. A. Williams, A. Spang, T. J. G. Ettema, Hikarchaeia demonstrate an intermediate stage in the methanogen-to-halophile transition. *Nat. Commun.* **11**, 5490 (2020).

62. A. B. Narrowe, A. Spang, C. W. Stairs, E. F. Caceres, B. J. Baker, C. S. Miller, T. J. G. Ettema, Complex evolutionary history of translation elongation factor 2 and diphthamide biosynthesis in Archaea and parabasalids. *Genome Biol. Evol.* **10**, 2380–2393 (2018).

63. P. S. Adam, G. Borrel, C. Brochier-Armanet, S. Gribaldo, The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. *ISME J.* **11**, 2407–2425 (2017).

64. M. Aouad, J.-P. Flandrois, F. Jauffrit, M. Gouy, S. Gribaldo, C. Brochier-Armanet, A divide-and-conquer phylogenomic approach based on character supermatrices resolves early steps in the evolution of the Archaea. *BMC Ecol. Evol.* **22**, 1 (2022).

65. M. Aouad, N. Taib, A. Oudart, M. Lecocq, M. Gouy, C. Brochier-Armanet, Extreme halophilic archaea derive from two distinct methanogen Class II lineages. *Mol. Phylogenet. Evol.* **127**, 46–54 (2018).

66. N. Dombrowski, J.-H. Lee, T. A. Williams, P. Offre, A. Spang, Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS Microbiol. Lett.* **366** (2019).

67. B. J. Baker, V. De Anda, K. W. Seitz, N. Dombrowski, A. E. Santoro, K. G. Lloyd, Diversity, ecology and evolution of Archaea. *Nat Microbiol* **5**, 887–900 (2020).

68. J. P. Beam, E. D. Becraft, J. M. Brown, F. Schulz, J. K. Jarett, O. Bezuidt, N. J. Poulton, K. Clark, P. F. Dunfield, N. V. Ravin, J. R. Spear, B. P. Hedlund, K. A. Kormas, S. M. Sievert, M. S. Elshahed, H. A. Barton, M. B. Stott, J. A. Eisen, D. P. Moser, T. C. Onstott, T. Woyke, R. Stepanauskas, Ancestral absence of electron transport chains in Patescibacteria and DPANN, *bioRxiv* (2020)p. 2020.04.07.029462.

69. C. Rinke, M. Chuvochina, A. J. Mussig, P.-A. Chaumeil, A. A. Davín, D. W. Waite, W. B. Whitman, D. H. Parks, P. Hugenholtz, A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nat Microbiol* **6**, 946–959 (2021).

70. K. Zaremba-Niedzwiedzka, E. F. Caceres, J. H. Saw, D. Bäckström, L. Juzokaite, E. Vancaester, K. W. Seitz, K. Anantharaman, P. Starnawski, K. U. Kjeldsen, M. B. Stott, T. Nunoura, J. F. Banfield, A. Schramm, B. J. Baker, A. Spang, T. J. G. Ettema, Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).

71. D. H. Parks, M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil, P. Hugenholtz, A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).

72. D. H. Parks, C. Rinke, M. Chuvochina, P.-A. Chaumeil, B. J. Woodcroft, P. N. Evans, P. Hugenholtz, G. W. Tyson, Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* **2**, 1533–1542 (2017).

73. D. Wacey, *Early Life on Earth: A Practical Guide* (Springer Science & Business Media, 2009).

74. M. Barboni, P. Boehnke, B. Keller, I. E. Kohl, B. Schoene, E. D. Young, K. D. McKeegan, Early formation of the Moon 4.51 billion years ago. *Sci Adv* **3**, e1602365 (2017).

75. B. B. Hanan, G. R. Tilton, 60025: relict of primitive lunar crust? *Earth Planet. Sci. Lett.* **84**, 15–21 (1987).

76. A. Criscuolo, S. Gribaldo, BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).

77. T. Kleine, H. Palme, K. Mezger, A. N. Halliday, Hf-W chronometry of lunar metals and the age and early differentiation of the Moon. *Science* **310**, 1671–1674 (2005).

78. N. Aghanim, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Banday, R. B. Barreiro, N. Bartolo, S. Basak, R. Battye, K. Benabed, J.-P. Bernard, M. Bersanelli, P. Bielewicz, J. J. Bock, J. R. Bond, J. Borrill, F. R. Bouchet, F. Boulanger, M. Bucher, C. Burigana, R. C. Butler, E. Calabrese, J.-F. Cardoso, J. Carron, A. Challinor, H. C. Chiang, J. Chluba, L. P. L. Colombo, C. Combet, D. Contreras, B. P. Crill, F. Cuttaia, P. de Bernardis, G. de Zotti, J. Delabrouille, J.-M. Delouis, E. Di Valentino, J. M. Diego, O. Doré, M. Douspis, A. Ducout, X. Dupac, S. Dusini, G. Efstathiou, F. Elsner, T. A. Enßlin, H. K. Eriksen, Y. Fantaye, M. Farhang, J. Fergusson, R. Fernandez-Cobos, F. Finelli, F. Forastieri, M. Frailis, A. A. Fraisse, E. Franceschi, A. Frolov, S. Galeotta, S. Galli, K. Ganga, R. T. Génova-Santos, M. Gerbino, T. Ghosh, J. González-Nuevo, K. M. Górski, S. Gratton, A. Gruppuso, J. E. Gudmundsson, J. Hamann, W. Handley, F. K. Hansen, D. Herranz, S. R. Hildebrandt, E. Hivon, Z. Huang, A. H. Jaffe, W. C. Jones, A. Karakci, E. Keihänen, R. Keskitalo, K. Kiiveri, J. Kim, T. S. Kisner, L. Knox, N. Krachmalnicoff, M. Kunz, H. Kurki-Suonio, G. Lagache, J.-M. Lamarre, A. Lasenby, M. Lattanzi, C. R. Lawrence, M. Le Jeune, P. Lemos, J. Lesgourgues, F. Levrier, A. Lewis, M. Liguori, P. B. Lilje, M. Lilley, V. Lindholm, M. López-Caniego, P. M. Lubin, Y.-Z. Ma, J. F. Macías-Pérez, G. Maggio, D. Maino, N. Mandolesi, A. Mangilli, A. Marcos-Caballero, M. Maris, P. G. Martin, M. Martinelli, E. Martínez-González, S. Matarrese, N. Mauri, J. D. McEwen, P. R. Meinhold, A. Melchiorri, A. Mennella, M. Migliaccio, M. Millea, S. Mitra, M.-A. Miville-Deschênes, D. Molinari, L. Montier, G. Morgante, A. Moss, P. Natoli, H. U. Nørgaard-Nielsen, L. Pagano, D. Paoletti, B. Partridge, G. Patanchon, H. V. Peiris, F. Perrotta, V. Pettorino, F. Piacentini, L. Polastri, G. Polenta, J.-L. Puget, J. P. Rachen, M. Reinecke, M. Remazeilles, A. Renzi, G. Rocha, C. Rosset, G. Roudier, J. A. Rubiño-Martín, B. Ruiz-Granados, L. Salvati, M. Sandri, M. Savelainen, D. Scott, E. P. S. Shellard, C. Sirignano, G. Sirri, L. D. Spencer, R. Sunyaev, A.-S. Suur-Uski, J. A. Tauber, D. Tavagnacco, M. Tenti, L. Toffolatti, M. Tomasi, T. Trombetti, L. Valenziano, J. Valiviita, B. Van Tent, L. Vibert, P. Vielva, F. Villa, N. Vittorio, B. D. Wandelt, I. K. Wehus, M. White, S. D. M. White, A. Zacchei, A. Zonca, Planck Collaboration, Planck 2018 results. VI. Cosmological parameters, *arXiv [astro-ph.CO]* (2018). http://arxiv.org/abs/1807.06209.

79. P. M. Shih, J. Hemp, L. M. Ward, N. J. Matzke, W. W. Fischer, Crown group Oxyphotobacteria postdate the rise of oxygen. *Geobiology* **15**, 19–29 (2017).

80. A. J. Drummond, S. Y. W. Ho, M. J. Phillips, A. Rambaut, Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).

81. A. M. Satkoski, N. J. Beukes, W. Li, B. L. Beard, C. M. Johnson, A redox-stratified ocean 3.2 billion years ago. *Earth Planet. Sci. Lett.* **430**, 43–53 (2015).

82. A. A. Davín, E. Tannier, T. A. Williams, B. Boussau, V. Daubin, G. J. Szöllősi, Gene transfers can date the tree of life. *Nat Ecol Evol* **2**, 904–909 (2018).

83. G. P. Fournier, K. R. Moore, L. T. Rangel, J. G. Payette, L. Momper, T. Bosak, The Archean origin of oxygenic photosynthesis and extant cyanobacterial lineages. *Proc. Biol. Sci.* **288**, 20210675 (2021).

84. G. J. Szöllősi, S. Höhna, T. A. Williams, D. Schrempf, V. Daubin, B. Boussau, Relative time constraints improve molecular dating. (2021). https://doi.org/10.1093/sysbio/syab084.

85. J. M. Wolfe, G. P. Fournier, Horizontal gene transfer constrains the timing of methanogen evolution. *Nat Ecol Evol* **2**, 897–903 (2018).

86. P. Kapli, T. Flouri, M. J. Telford, Systematic errors in phylogenetic trees. *Curr. Biol.* **31**, R59–R64 (2021).

87. P. Kapli, Z. Yang, M. J. Telford, Phylogenetic tree building in the genomic age. *Nat. Rev. Genet.* **21**, 428–444 (2020).

88. P. Lopez, P. Forterre, H. Philippe, The root of the tree of life in the light of the covarion model. *J. Mol. Evol.* **49**, 496–508 (1999).

89. B. Morel, P. Schade, S. Lutteropp, T. A. Williams, G. J. Szöllősi, A. Stamatakis, SpeciesRax: A Tool for Maximum Likelihood Species Tree Inference from Gene Family Trees under Duplication, Transfer, and Loss. *Mol. Biol. Evol.* **39**, msab365 (2022).

90. G. J. Szöllősi, W. Rosikiewicz, B. Boussau, E. Tannier, V. Daubin, Efficient exploration of the space of reconciled gene trees. *Syst. Biol.* **62**, 901–912 (2013).

91. C. Zhang, C. Scornavacca, E. K. Molloy, S. Mirarab, ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy. *Mol. Biol. Evol.* **37**, 3292–3307 (2020).

92. Q. Zhu, *WoL: A Reference Phylogeny for Bacterial and Archaeal Genomes* (GitHub, 2022; https://github.com/biocore/wol).

93. T. Seemann, Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

94. T. Aramaki, R. Blanc-Mathieu, H. Endo, K. Ohkubo, M. Kanehisa, S. Goto, H. Ogata, KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).

95. A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Studholme, C. Yeats, S. R. Eddy, The Pfam protein families database. *Nucleic Acids Res.* **32**, D138-41 (2004).

96. D. H. Haft, J. D. Selengut, O. White, The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373 (2003).

97. B. L. Cantarel, P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, B. Henrissat, The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* **37**, D233-8 (2009).

98. N. D. Rawlings, A. J. Barrett, R. Finn, Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* **44**, D343-50 (2016).

99. M. H. Saier Jr, C. V. Tran, R. D. Barabote, TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res.* **34**, D181-6 (2006).

100. D. Søndergaard, C. N. S. Pedersen, C. Greening, HydDB: A web tool for hydrogenase classification and analysis. *Sci. Rep.* **6**, 34212 (2016).

101. P. Jones, D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.-Y. Yong, R. Lopez, S. Hunter, InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

102. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

103. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29-37 (2011).

104. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

105. N. Dombrowski, *Phylogeny Tutorial* (Github, 2022; https://github.com/ndombrowski/Phylogeny_tutorial/tree/main/Input_files/5_required_scripts).

106. K. Katoh, H. Toh, Recent developments in the MAFFT multiple sequence alignment program. (2008). https://doi.org/10.1093/bib/bbn013.

107. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

108. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

109. Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

110. P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, D. H. Parks, GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).

111. R Development Core Team, *R: A Language and Environment for Statistical Computing* (2021; https://www.R-project.org).

112. H. Wickham, *Ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).

CHAPTER 3

ATP synthase evolution on a cross-braced dated tree of life

Tara A. Mahendrarajah, Edmund R.R. Moody*, Dominik Schrempf*, Lénárd Szánthó, Nina Dombrowski, Adrián A. Davín, Davide Pisani, Philip C.J. Donoghue, Gergely J. Szöllősi, Tom A. Williams & Anja Spang
*these authors contributed equally to this work

Nature Communications, 2023

SUMMARY AND CONTRIBUTIONS

The primary goal of this project was to address the evolutionary history of the ATP synthase, considering the timing of major events in cellular evolution. Timing deep evolutionary events is challenging, in part due to the paucity of prokaryotic fossils and the difficulty of modeling rate variation using current molecular clock methods. To overcome this, we added eukaryotes with available fossil evidence to our phylogenetic analysis, which provide a lens into deep prokaryotic evolution due to the endosymbioses involved in eukaryogenesis. The phylogenetic assessment centered on marker genes that are shared between the Archaea, Bacteria, and eukaryotes. Eukaryotes have inherited up to three distinct sets of ribosomes and ATP synthases from their prokaryotic ancestors – a nuclear source from their ancestral archaeal host, a mitochondrial source from the alphaproteobacterial ancestor of the mitochondria, and a plastid source from the cyanobacterial ancestor of the plastid. The core site of ATP synthesis in the ATP synthase consists of two components, the so-called catalytic and non-catalytic subunits, which were derived from an ancient pre-LUCA gene duplication followed by a loss of function in one subunit. As such, speciation events appear twice in the ATP synthase gene tree. Therefore, ancient gene duplications (ATP synthase) and endosymbioses (ribosomes and ATP synthase) result in node equivalency across their respective phylogenies, more specifically, multiple nodes on the same tree correspond to the same evolutionary event. We used this principle to develop a cross-bracing approach, where equivalent nodes are "braced" together to better constrain fossil calibrations that were applied in our molecular dating analysis. Here we combined cross-bracing with a relative time constraint (*1*), to propagate eukaryotic fossil information into poorly modeled regions of the tree of life (TOL), namely the deep prokaryotic branches.

When we applied this technique to the ATP synthase gene tree, we found that the split between the catalytic and non-catalytic subunits occurred very early in the evolutionary timeline between 4.52-4.46 Ga, which predates or at minimum, co-occurred with the last universal common ancestor (LUCA; 4.52-4.32 Ga) and the last bacterial common ancestor (LBCA, 4.49-4.05 Ga). The divergence between the F- and A/V-type ATP synthases was also resolved to a similar period, between 4.52-4.38 Ga in the catalytic clade and 4.52-4.42 Ga in the non-catalytic clade. Interestingly, the last archaeal common ancestor (LACA) was inferred to 3.95-3.37 Ga, either implying a sampling bottleneck or suggesting an alternative scenario of ATP synthase evolution that does not parallel the divergence between the Archaea and Bacteria from LUCA. In separate analyses we reconciled the catalytic and non-catalytic subunits of both F- and A/V-type ATP synthases to LUCA, found that bacterial A/V-type ATP synthases are widespread, and that F-type archaeal ATP synthases are restricted to a select few Methanosarcina. Considering these data together, we proposed two alternative scenarios for ATP synthase evolution: 1) that an ancestral F-type-like rudimentary ATP synthase existed in LUCA, evolved along the branches to LBCA and LACA, with a transfer from LACA to the stem toward LBCA, or 2) both the F- and A/V-type ATP synthases already existed in LUCA, and evolved along each branch to LBCA and LACA with a loss of the F-type along the archaeal stem. Either scenario is consistent

with other evidence indicating that LBCA contains both the F- and A/V-type ATP synthases (see Chapter 4) and the ubiquity of the A/V-type bacterial complexes across the TOL. Altogether, results provide greater insight into ATP synthase evolution and a timeline for cellular evolution.

This was a collaborative project that required the integration of expertise in phylogenetics, phylogenomics, gene tree-species tree reconciliation, molecular dating, and fossil calibration application, among others. Anja Spang and I conceptualized this project and the methodological approaches. Briefly, as the first author, I contributed the majority of analyses toward the completion of this study. I manually curated the taxonomic dataset including 350 Archaea, 350 Bacteria, and 100 Eukaryota. I performed the majority of the phylogenetic and phylogenomic analyses and spent a considerable amount of time testing methods, including gene tree filtering, different alignment trimming/filtering procedures, model testing, topology testing, constraint trees, different taxonomic sets, among others. Reconciliations of the catalytic and non-catalytic subunits of the F- and A/V-type ATP synthases was performed by Edmund R.R. Moody while I performed ancestral sequence reconstructions. Software for the cross-bracing molecular dating procedure was developed and executed by Dominik Schrempf and Tom A. Williams and I formulated the cross-bracing, fossil calibration, and relative constraint input files that assigned all nodes (equivalent and relative) to the appropriate fossil ages with the advice of Philip C.J. Donoghue and David Pisani. All authors contributed methods or developed scripts for figures and/or computational analysis. I wrote the first draft and myself, Anja Spang, Tom A. Williams, and Edmund R.R. Moody conducted weekly meetings to write the final draft of the manuscript.

References

1. G. J. Szöllősi, S. Höhna, T. A. Williams, D. Schrempf, V. Daubin, B. Boussau, Relative Time Constraints Improve Molecular Dating. *Syst. Biol.* **71**, 797–809 (2022).

ABSTRACT

The timing of early cellular evolution, from the divergence of Archaea and Bacteria to the origin of eukaryotes, is poorly constrained. The ATP synthase complex is thought to have originated prior to the Last Universal Common Ancestor (LUCA) and analyses of ATP synthase genes, together with ribosomes, have played a key role in inferring and rooting the tree of life. We reconstruct the evolutionary history of ATP synthases using an expanded taxon sampling set and develop a phylogenetic cross-bracing approach, constraining equivalent speciation nodes to be contemporaneous, based on the phylogenetic imprint of endosymbioses and ancient gene duplications. This approach results in a highly resolved, dated species tree and establishes an absolute timeline for ATP synthase evolution. Our analyses show that the divergence of ATP synthase into F- and A/V-type lineages was a very early event in cellular evolution dating back to more than 4 Ga, potentially predating the diversification of Archaea and Bacteria. Our cross-braced, dated tree of life also provides insight into more recent evolutionary transitions including eukaryogenesis, showing that the eukaryotic nuclear and mitochondrial lineages diverged from their closest archaeal (2.67-2.19 Ga) and bacterial (2.58-2.12 Ga) relatives at approximately the same time, with a slightly longer nuclear stem-lineage.

INTRODUCTION

The phylogeny and timeline of early cellular evolution, including the age of the last universal common ancestor (LUCA), the radiations of the archaeal and bacterial domains, and the origin of eukaryotes and their progenitor prokaryote lineages, is poorly constrained (*1*). Recent genomics approaches have greatly improved our sampling of natural diversity and uncovered previously unknown microbial lineages that are key to understanding early cellular evolution (*2*). For instance, the Asgard archaea (*3–5*) (also referred to as Asgardarchaeota (*6*), Supplementary Data 1), include the closest known sister lineage of the Eukaryota (i.e. eukaryotes) (*3*, *4*, *7*, *8*) and have provided support for the evolution of the eukaryotic cell through a symbiosis between at least one asgardarchaeal and one alphaproteobacterial partner (*9–15*). The discovery of the symbiotic and genome-reduced members of the bacterial Candidate Phyla Radiation (CPR) and the DPANN archaea (named after the first member lineages of this group, the Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeaota, and Nanohaloarchaeota) (*16–18*), that were originally interpreted as early diverging branches on each side of the root of the tree of life (*2*), might be important for our understanding of the deep split separating Archaea and Bacteria (*19*). However, more recent phylogenomic analyses suggest that CPR are instead sister to Chloroflexota within Terrabacteria (*19–23*). Time scaling molecular evolution is challenging because the rate of molecular evolution has varied substantially through time (*19*, *24–27*) and, with few fossil calibrations (e.g. maximum age constraints and lack of Precambrian maximum age calibrations), clock models struggle to capture this rate variation. This has led to estimates of divergence time that in some cases are uncertain (as in the case of the age of LUCA – 4.52–4.48 Ga (*19*, *27*, *28*)). Additional sources of temporal information beyond fossil and geochemical calibrations are crucial to improve these estimates of divergence time.

The ATP synthase is a protein complex central to energy conservation through the synthesis and hydrolysis of ATP (*29*, *30*). It is a useful marker to address key evolutionary transitions due to the presence of this enzyme across all domains of life (*24*, *29*, *31–37*). The ATP synthase family is classified into the F-, A-, and V-type ATP synthases (*30*, *33*, *34*) based on taxonomic affiliation, function, and cellular localization (*30*, *34*, *38*). F-type ATP synthases are ubiquitous across bacteria and eukaryotes and localize to cellular, mitochondrial, and plastid membranes (*39*). In line with this, eukaryotic F-type ATP synthases are hypothesized to be derived from the bacterial ancestors of these organelles (*33*, *35*, *40*). The A-type ATP synthase (*34*, *38*, *41*), found primarily in Archaea, belongs to a larger family of A/V-type ATP synthases (*38*) that also include eukaryotic complexes found in vacuoles (*32–34*, *38*, *40–42*). The F- and A/V-type ATP synthases share a common foundational architecture consisting of a soluble cytoplasmic component (R1) connected to an insoluble membrane component (R0) (Supplementary Fig. 1). The hexameric headpiece of the R1 complex contains three copies each of a catalytic (*c*) and non-catalytic (*nc*) subunit and is the site of ATP synthesis and hydrolysis. The catalytic and non-catalytic subunits comprising the soluble hetero-hexameric R1 component, are paralogs to each other. They arose prior to LUCA through an ancient duplication of a RecA family protein (P-loop NTPase)

followed by the loss of the catalytic function in one subunit (*29*, *32–35*, *38*, *40*, *43*). Due to this ancient gene duplication, each paralog can act as an outgroup to the other, providing a way to root the tree of life (*44*). Since the duplication occurred before the divergence between Archaea and Bacteria, speciation events during the subsequent history of life appear at least twice in the ATP synthase gene tree. This circumstance has been used to improve date estimates for eukaryotic evolution by "cross-bracing" (constraining to the same unknown age) equivalent speciation nodes in the gene tree (*24*), which propagates limited fossil evidence across the tree. Cross-bracing improves clock estimates in two ways. First, the information that two distant nodes in the tree must have the same age provides a useful constraint on the rates of evolution and ages of the intervening branches. Second, it enables the fossil record of eukaryotes to inform divergence times within both the archaeal and bacterial domains, because eukaryotes obtained ATP synthase paralogs from both sources. In principle, this approach might be expanded beyond ATP synthase to the core set of ribosomal proteins that are conserved between the nucleus, mitochondrion, and chloroplast of eukaryotes as a result of the mitochondrial and plastid endosymbioses. Cross-bracing a ribosomal species tree could help to avoid difficulties arising from HGT events during ATP synthase evolution (*32*, *33*, *40*, *45*), and the limited resolving power of single gene trees.

To improve our understanding of the evolutionary history of ATP synthases and cellular evolution, we perform phylogenetic analyses using an updated taxon sampling set, ancestral sequence reconstruction (*46*, *47*), and novel molecular dating approaches including cross-bracing (*24*, *48*, *49*). We also use probabilistic gene- and species-tree reconciliation methods (implemented in Amalgamated Likelihood Estimation (ALE)) (*50–52*) to determine the origin and evolution of the ATP synthase and each of its subfamilies. ALE allows us to compare the ATP synthase gene family tree to the tree of life and to infer the history of gene duplication, transfer, and loss during ATP synthase evolution. We assemble a set of ribosomal marker proteins that includes three distinct clades of eukaryotic homologs derived from archaeal, alphaproteobacterial, and cyanobacterial ancestors. Due to gene duplications (ATP synthase) and endosymbiosis events (ribosomal marker genes and ATP synthase), cross-bracing can be applied to both datasets. Our analyses confirm the split of the catalytic and non-catalytic ATP synthase subunits prior to LUCA and reveal the prevalence and early evolution of A/V-type ATP synthases in Bacteria. Our dating analyses establish absolute age estimates for LUCA, the Last Bacterial Common Ancestor (LBCA), and the Last Archaeal Common Ancestor (LACA). We link early cellular evolution with the origin of the head component of ATP synthases, which has diversified earlier than previously assumed. Finally, our analyses improve time estimates for the origin of eukaryotes from its prokaryotic ancestors and thereby inform on eukaryogenesis.

RESULTS

DISTRIBUTION OF ATP SYNTHASES ACROSS BACTERIA, ARCHAEA AND EUKARYOTES

We analysed the distribution of ATP synthase genes across our reference genome dataset of 800 Archaea, Bacteria, and eukaryotes (Figs. 1–2, Supplementary Fig. 2, Supplementary Data 1–4). In agreement with previous work (*29*, *32–36*, *38*, *40*), our results indicate a partitioning of the F- and A/V-type ATP synthases by domain. Archaea and Bacteria contain primarily A/V-type and F-type subunits, respectively, and eukaryotes harbor complexes of both types (Fig. 1, Supplementary Fig. 2). However, in Bacteria the pattern is more complex than generally assumed (*33*, *35*, *53*). Consistent with emerging evidence that several bacteria contain A/V-type ATP synthases (*35*), we found that 46% (23/50) of bacterial phylum-level lineages encode genes for A/V-type ATP synthases in conjunction with (n = 19), or to the exclusion of (n = 4), a bacterial F-type ATP synthase (Fig. 1, Supplementary Fig. 2, Supplementary Data 4). Conversely, only three members of a single archaeal lineage, Methanosarcinales, contain F-type ATP synthases in addition to their A/V-type complex (Fig. 1, Supplementary Fig. 2, Supplementary Data 4), as observed previously (*54–56*).

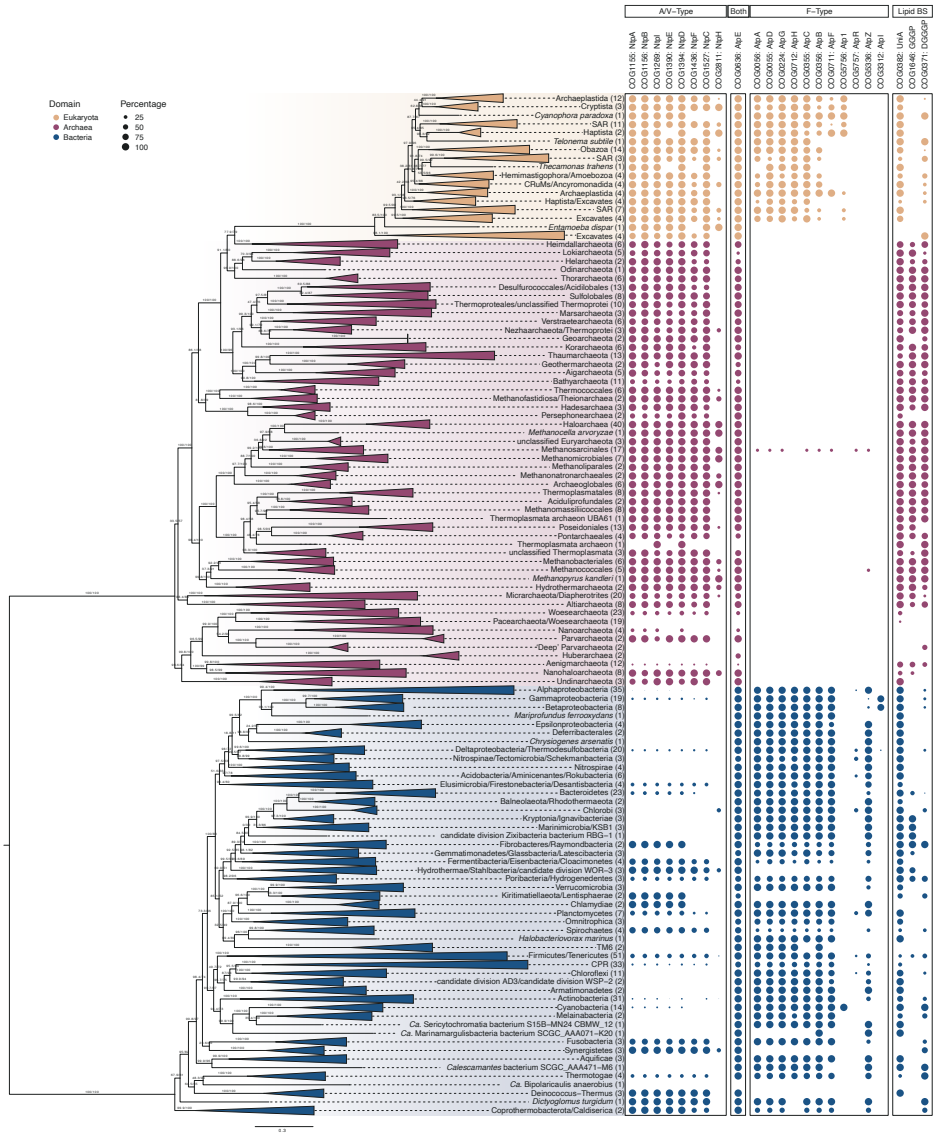


Fig. 1 | Distribution of COG families representing the F- and A/V-type ATP synthase subunits and select lipid biosynthesis genes across the tree of life. COG families corresponding to the ATP synthase subunits and lipid biosynthesis genes (see Methods for selection of COG families, Supplementary Data 3) are represented as a percentage presence by phylogenetic cluster, consistent with collapsed taxonomic clades in the maximum-likelihood concatenated species tree. The concatenated alignment contains 780 taxa and was trimmed with BMGE v1.12 (settings: -m BLOSUM30 -h 0.55) *(110)* to remove poorly-aligning positions (final alignment length = 3367 amino acids). The maximum-likelihood tree was inferred using IQ-TREE2 v2.1.2 with the LG+C20+R+F model with SH-like approximate likelihood (left) and ultrafast bootstrap approximation (right), each with 1000 replicates *(111, 122, 123)*. The scale bar corresponds to the expected number of substitutions per site. Color code: archaea = red, bacteria = blue, eukaryotes = yellow.

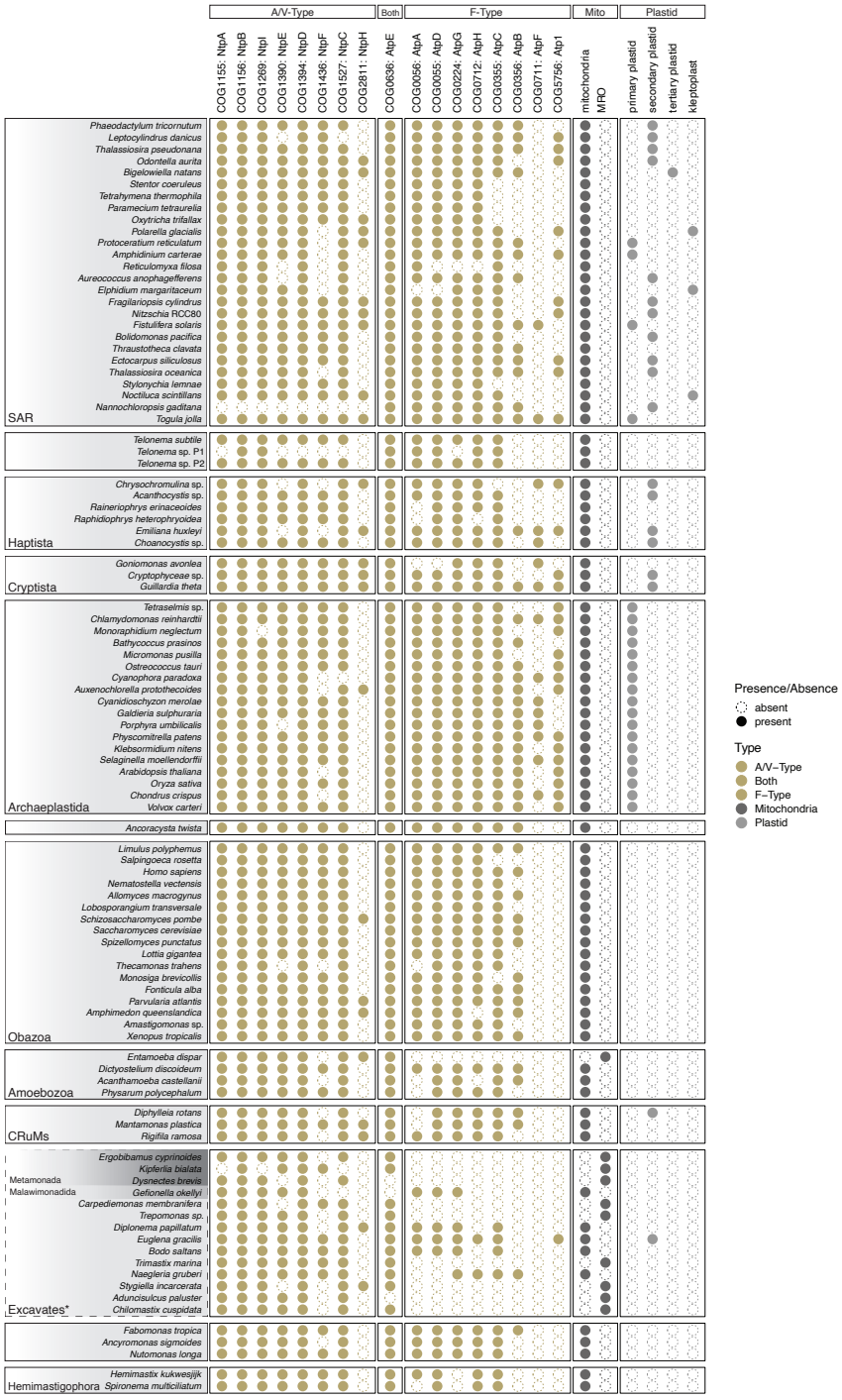


Fig. 2 | Occurrence of COG families representing the F- and A/V-type ATP synthase subunits and the presence/absence of key metabolic organelles across the 100 sampled eukaryotes. COG families

representing the ATP synthase subunits (see Methods for selection of COG families, Supplementary Data 3) are presented as binary presence-absence counts per taxon. The relationships among eukaryotic supergroups is consistent with Burki, 2020 (*136*). Dashed lines represent groups with greater uncertainty. Mito = mitochondria and mitochondrion-related organelles (MROs). Plastid = primary-, secondary-, and tertiary-plastid, and kleptoplast. See Supplementary Data 5 for additional information on organelle distribution. The list of eukaryotic ATP synthase sequences flagged as putative bacterial contamination can be found in Supplementary Data 4.

Despite the core role of ATP synthases in energy conservation, some prokaryotes (including members of the DPANN archaea (*57*–*59*)) lack functional homologs (Fig. 1, Supplementary Fig. 2, Supplementary Data 4). These absences are present across related lineages, suggesting a genuine loss, rather than metagenome-assembled genome incompleteness. Other DPANN lineages, such as Nanohaloarchaeota, may have inherited their ATP synthase from a DPANN ancestor (Supplementary Fig. 6) or acquired ATP synthase genes from symbiotic partners (Supplementary Fig. 7–10, Discussion) (*45*, *60*). Several Bathyarchaeota lack ATP synthase complexes (Fig. 1, Supplementary Fig. 2, Supplementary Data 4), consistent with the previously noted absence of ATP synthases in the Bathyarchaeote BA1 and BA2 (*61*, *62*). Instead, these organisms may produce ATP through substrate-level phosphorylation using a putative ATP-forming acetyl-CoA synthetase (*61*, *62*). We found that 60.6% (20/33) of the analysed CPR encode F-type ATP synthases, with phylogenetic trees suggesting inheritance from a common ancestor with Chloroflexi (Fig. 1, Supplementary Fig. 2, Supplementary Data 4). Conversely, 30.3% (10/33) of the sampled CPR have lost genes that would enable the formation of a canonical ATP synthase complex (Supplementary Fig. 2, Supplementary Data 4). Interestingly, three members of the CPR in our dataset lack an F-type ATP synthase but have a near-complete or complete A/V-type complex (Fig. 1, Supplementary Fig. 2), indicating a recent acquisition of the A/V-type complex via HGT potentially by members of the Synergistetes (Supplementary Fig. 2, 6–10).

Most eukaryotic lineages contained core functional subunits of both F- and A/V-type ATP synthases, with the exception of 24/100 analyzed representatives, including *Entamoeba dispar* and certain Excavates (Fig. 2, Supplementary Data 5). This is consistent with the energy metabolism of these anaerobes whose mitochondrion-related organelles have lost components of the aerobic electron transport chain (*63*, *64*) (Fig. 2). We observed that 78% (14/18) of Archaeplastida encode genes for Atp1, the F-type alpha subunit of cyanobacteria (COG5756, Fig. 2, Supplementary Data 5). Notably, this gene is lacking in species without a plastid (Fig. 2), consistent with the existence of a second F-type ATP synthase of endosymbiotic origin in chloroplasts. However, 36% of plastid-bearing eukaryotes (16/45) lack an Atp1 homolog (Fig. 2), indicating subsequent loss of Atp1 in some photosynthetic eukaryotes.

EVOLUTIONARY HISTORY OF SOLUBLE ATP SYNTHASE SUBUNITS

Phylogenetic analyses including all catalytic (*c*) and non-catalytic (*nc*) subunits of the soluble head component of the F- and A/V-type ATP synthase (Supplementary Fig. 1), the F1 beta and A1/V1A and the F1 alpha and A1/V1B, respectively (hereafter referred to as *c*F1, *c*A1V1, *nc*F1, *nc*A1V1, revealed four clades corresponding to each of the four protein families (Fig. 3A, Supplementary Fig. 10). Based on the gene family tree and the observation that all organisms encoding an ATP

synthase possess catalytic (*c*F1 and *c*A1V1) (Fig. 3A, Supplementary Figs. 4, 6, 8–10) and non-catalytic (*nc*F1 and *nc*A1V1) (Fig. 3A, Supplementary Figs. 5, 7–10) subunits, our analysis agrees with the consensus view (*31*, *32*, *35*, *40*), that the deepest split lies between those families (Fig. 3A, Supplementary Fig. 10) (*31*, *32*, *35*, *40*). Our results suggest an early divergence of the functional capacities of each subunit followed by subsequent bifurcations into F- and A/V-type complexes (Fig. 3A, Supplementary Fig. 10). The deep splits observed within each of the catalytic and non-catalytic subunits of the F- and A/V-type complexes have been hypothesized to coincide with the speciation of Archaea and Bacteria (*31*, *32*, *65*) (Fig. 3A).

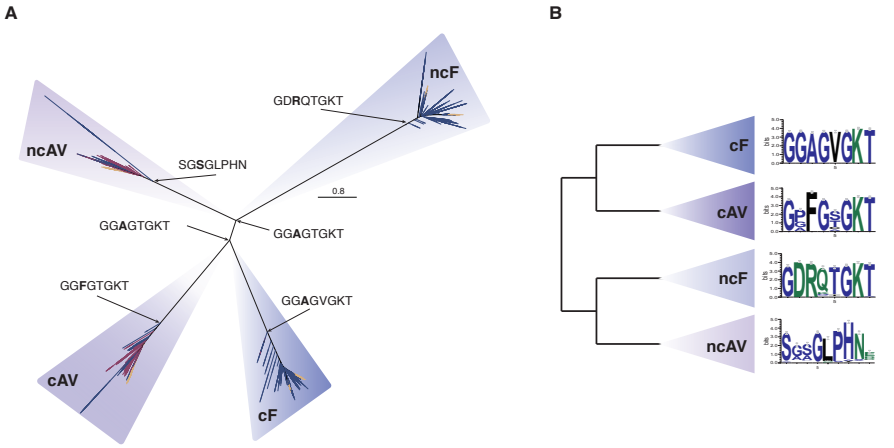


Fig. 3 | Maximum-likelihood tree of all ATP synthase headpiece subunits identified in sampled Archaea (red), Bacteria (blue), and Eukaryotes (yellow). A Homologs corresponding to each subunit form monophyletic clusters for each protein family. Catalytic subunits (*c*F1 and *c*A1V1) and non-catalytic subunits (*nc*F1 and *nc*A1V1) cluster together on either side of the root. The alignment contains 1520 sequences and was trimmed with BMGE v1.12 (settings: -m BLOSUM30 -h 0.55) *(110)* (alignment length = 350 amino acids). The maximum-likelihood tree was inferred using IQ-TREE2 v2.1.2 with the LG+C50+R+F model, selected using the best-fitting model (chosen by BIC) *(111, 122, 132)*. The scale bar corresponds to the expected number of substitutions per site. The Walker-A motif from ancestrally reconstructed sequences *(123)* are shown at their respective nodes. **B** Conserved protein motifs for each subunit derived from the same alignment.

Determining which of the four head-forming subunits was present in LACA, LBCA, and LUCA based on gene tree inspection is challenging. For instance, the identification of A/V-type ATP synthases in many Bacteria (Fig. 1, Supplementary Fig. 2) and the recent inference of the presence of components of both F- and A/V-Type ATP synthases in the genome of LBCA (*21*), challenge a late horizontal acquisition of the A/V-type ATP synthase by Bacteria. To evaluate these hypotheses within a statistical framework, we used the ALE probabilistic approach (*51*) to reconcile gene trees for each of the ATP synthase subunits with the species tree as a whole, using distinct data treatments (Methods, Supplementary Data 6). This approach compares the gene family tree with the species tree to infer gene origination, duplication, transfer, and loss events. It maps branches of the gene tree to the species tree, using conditional clade probabilities (*66*) to account for uncertainty in the gene family tree analyses (*51*). These analyses agreed

with our manual inspection of the gene trees, suggesting that the *c*A1V1 and *nc*A1V1 subunits were present in LACA (presence probability, PPs = 0.99–1) (*67*) and the *c*F1 and *nc*F1 subunits were present in LBCA (PPs = 0.99–1) (*21*). We recovered support for the presence of the *c*A1V1 (PPs = 0.64–1) subunit in LBCA, as has been suggested recently (*21*), while the presence of the *nc*A1V1 subunit in LBCA was supported only in trees inferred using the LG+C20+R+F but not LG+C60+R+F model (C20: PPs = 0.99–1, C60: PPs = 0.21–0.28, Supplementary Data 6).

The *nc*F1 (PPs = 0.79–1), *nc*A1V1 (PPs = 0.99–1), *c*F1 (PPs = 1), and *c*A1V1 (PPs = 0.99–1) gene families were estimated to having been present in LUCA, suggesting a putative pre-LUCA duplication of both the catalytic and non-catalytic subunits into the F- and A/V-type lineages (Supplementary Data 6, Fig. 4). However, deep branches in the gene trees are susceptible to systematic error, and distinguishing ancestral presence from early horizontal acquisition is difficult (*21*). Nonetheless, the widespread presence of genes encoding A/V-type subunits in modern Bacteria (Fig. 1, Supplementary Fig. 2, Supplementary Data 4) suggests that these genes were acquired early in bacterial evolution.

The presence of all four subunits in Bacteria is consistent with ideas for a root of the universal tree within Bacteria (*68–70*). However, we obtained significantly lower gene family likelihoods (approximately unbiased (AU) test): C60 model, pAU = 0.00009; C20 model, pAU = 0.0002) (Supplementary Data 6) for ATP synthase subunits reconciled with a species tree rooted within Bacteria rather than between Archaea and Bacteria (*20*, *21*). When eukaryotes were excluded, the within-Bacteria root also had a lower likelihood, though not significant (C60, pAU = 0.093; C20, pAU = 0.547) (Supplementary Data 6). These results agree with the consensus root between Archaea and Bacteria.

To investigate the key motifs characterizing the catalytic and non-catalytic subunits of the F- and A/V-type ATP synthases we examined conserved protein motifs in extant taxa. We focused on the sequence identity of the Walker-A motif (Supplementary Discussion), which has an amino acid composition of *GXXXXGKT* (*43*). This region comprises the primary "P-loop" domain responsible for binding phosphate during ATP synthesis/hydrolysis and is highly conserved across phosphate-binding proteins and fundamental to the activity of the ATP synthase (*71*, *72*). Our analyses of the Walker-A motifs across the *nc*F1, *c*F1, and *c*A1V1 subunits revealed a conserved motif with variation in positions 2–5 (Fig. 3B). However, the *nc*A1V1 subunit lacks a recognizable Walker-A motif and instead contains a *SGSGLPHN* motif in the corresponding position (Fig. 3B). The phosphate binding properties of this motif are unknown (*73*).

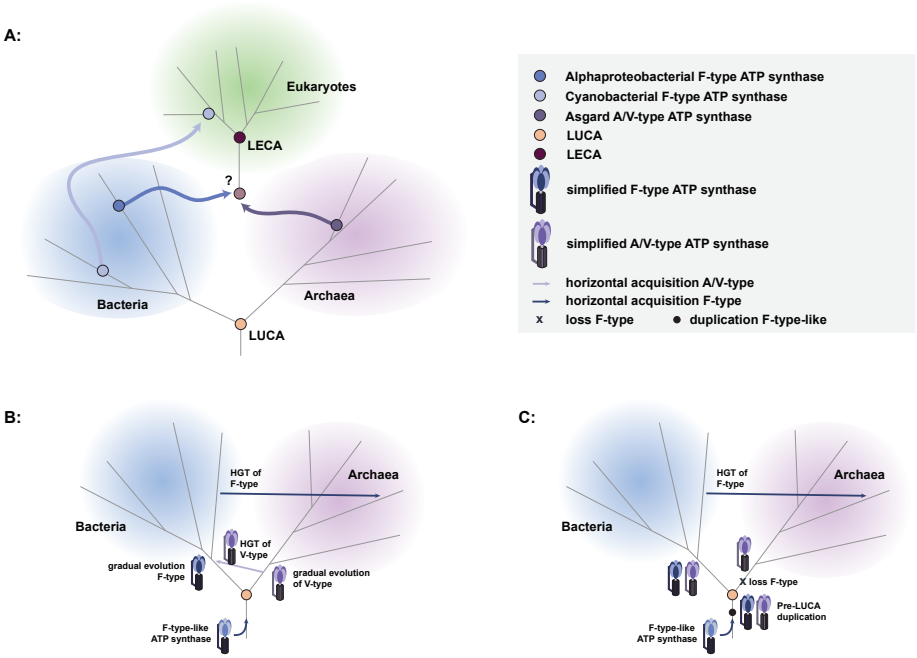


Fig. 4 | ATP synthase evolutionary scenarios. A Overview of possible ancestral ATP synthase acquisition in LECA from the putative prokaryotic hosts; the A/V-type derived from the archaeal host, and F-type ATP synthases derived from bacterial endosymbionts. **B** Evolutionary proposal supporting an F-type-like ancestral ATP synthase present pre-LUCA with subsequent divergence consistent with the split between Bacteria and Archaea and early transfers of A/V-type ATP synthases into the bacterial stem, and late HGT of F-type ATP synthases to Archaea. **C** Evolutionary proposal supporting an F-type-like ancestral ATP synthase and pre-LUCA duplication and divergence of the at least the head components of the F- and A/V-type subunits with subsequent loss of the F-type components along the archaeal stem. The cartoon of the ATP synthase was drawn manually in Adobe illustrator.

We performed ancestral sequence reconstruction (*46*, *47*) on the alignment of the unrooted combined phylogeny (Fig. 3A, Supplementary Data 7) to determine the ancestral sequence at the root of each of the four subunits (*nc*F1: Node123; *c*F1: Node126; *c*A1V1: Node516; and *nc*A1V1: Node883) as well as the root of the catalytic versus non-catalytic subunits (Fig. 3A) (*nc*F1 and *nc*A1V1: Node124; *c*F1 and *c*A1V1: Node125). Consistent with our observations of the conserved extant motifs, we found Walker-A motifs in the reconstructed sequences for the *nc*F1, *c*F1, and *c*A1V1 families and the alternative motif (*SGSGLPHN*) for the *nc*A1V1 family (Fig. 3A). The alanine (A) and phenylalanine (F) dichotomy in the third position of the *nc*F1, *c*F1 and *c*AV ancestors is consistent with previous findings distinguishing F- and A/V-type ATP synthase catalytic binding loops, respectively (Fig. 3A) (*41*). A motif pattern of *GGAGTGKT* was inferred for both the ancestor of the catalytic and non-catalytic subunits (Fig. 3A). This is compatible with a previously proposed scenario in which the progenitor ATP synthase was suggested to have contained six catalytic sites similar to the *c*F1 (*34*). While ancestral sequences inferred for the *nc*F1, *nc*A1V1, *c*F1 and *c*A1V1 are most similar to those in extant representatives of each of those families, the sequences inferred for each ancestor of the *nc* and *c* families

were both most similar to extant members of the *c*F1 subunits from F-type ATP synthases (Fig. 3, Supplementary Data 7). Taken together, this may indicate that the ancestral head component of the ATP synthase was more similar to the R1 complex of F-type ATP synthase and is consistent with the hypothesis that they evolved by duplication from a catalytic ancestor belonging to the "P-loop" NTPases (*24*, *29*, *31*, *34*, *37*). Furthermore, our results imply that the ncA1V1 subunit lost its Walker-A motif after divergence from the other subunits, though the functional consequence of the degenerated binding loop in the ncA1V1 subunit is unknown.

THE ORIGINS OF ATP SYNTHASES IN EUKARYOTES

In agreement with symbiogenetic models for the origin of the eukaryotic cell (*9–15*), our ATP synthase phylogenies suggest that eukaryotes inherited A/V- and F-type ATP synthases from their archaeal and bacterial ancestors (Fig. 3A, Supplementary Figs. 4–10) (*11*, *32–34*, *40*). Specifically, the relationship between Asgard archaea and eukaryotes was evident in phylogenies of the *nc*A1V1 subunit (Supplementary Figs. 6–10, Supplementary Data 8), with the strongest bootstrap support being 95.8/95 (Supplementary Fig. 7, Supplementary Data 8). In phylogenies for the catalytic subunits, the position of the eukaryotic branch was mostly unresolved (Supplementary Discussion, Supplementary Figs. 6–10, and Supplementary Data 8). This might be due to selective constraints or functional divergence considering that the eukaryotic V-type ATP synthase has evolved to couple proton transport to ATP hydrolysis rather than functioning as ATP synthase (*32*, *34*, *42*). The origin of eukaryotic F-type sequences from Alphaproteobacteria and Cyanobacteria was consistently recovered across a range of analyses including both Bayesian and maximum-likelihood inferences (Supplementary Discussion, Supplementary Figs. 4–5, Supplementary Fig. 11, Supplementary Data 8, Zenodo data repository: https://doi.org/10.5281/zenodo.10012837 (*74*)). Within the F-type subunits, the *nc*F1 phylogenies placed the sequences of eukaryotic plastids sister to *Gloeomargarita lithophora*, the closest living relative of the plastid (*75*), while the *c*F1 phylogeny grouped plastids together with most Cyanobacteria (Supplementary Fig. 11).

DATING THE SPECIES TREE AND ESTABLISHING AN ABSOLUTE TIMELINE FOR ATP SYNTHASE EVOLUTION

To establish a timeframe for the evolution of the ATP synthase, we built on the approach of Shih and Matzke (2013) (*24*) by bracing equivalent speciation nodes in both the ATP synthase phylogeny and a universal species tree. We took advantage of the greatly expanded sampling of organisms sequenced since the previous study (1520 total *nc* and *c* ATP synthase subunit sequences included in this study versus 149 total sequences (*24*)) and applied more fossil calibrations (ATP synthase gene tree $n = 10$, species tree $n = 12$ versus $n = 7$ (*24*)) (Supplementary Discussion, Supplementary Data 9). We developed a new molecular dating software (McmcDate) that implements both cross-bracing (two nodes constrained to the same age) as well as relative age constraints (one node is constrained to always be younger than another node, as informed, for instance, based on horizontal gene transfer between donor and recipient lineages (*49*, *76*)) (Supplementary Discussion, Supplementary Data 9).

Molecular dating analyses revealed that bracing the nuclear, mitochondrial, and plastid eukaryotic clades had a significant overall impact (*Z*-test statistic was −233.0 with a *p*-value of 0.0) on inferred rates of evolution, which are 16.5% higher overall than in the non-braced analysis (2.6e-4 average number of substitutions per million years and site, for ribosomal protein tree; Fig. 5A, C, Supplementary Figs. 12–17, Supplementary Data 10). As a result, age ranges (measured as 95% highest posterior density: the boundaries of the central 95% highest posterior densities of the distributions on ages) are modestly younger in the braced analysis, though similar overall (Supplementary Data 10). We estimate that LUCA lived 4.52–4.32 Ga and 4.52–4.42 Ga in the braced and non-braced analysis, respectively (Fig. 5A, Supplementary Fig. 16). Ages towards the younger end of the spectrum from our braced analysis seem more plausible considering the Moon-forming impact at 4.52 Ga, though both ages imply a rapid origin of LUCA following this putative sterilization event. In the following, we focus on dates from the cross-braced analysis but corresponding age ranges without bracing can be found in Fig. 5 and Supplementary Data 10). Of the two prokaryotic domains, LBCA was inferred to be older than LACA (4.49–4.05 Ga versus 3.95–3.37 Ga) indicating higher extinction or lower sampling rates for the archaeal stem-lineage (Fig. 5A, Supplementary Fig. 16). Our analyses suggest that eukaryotes diverged from their closest known asgardarchaeal relatives 2.67–2.19 Ga (Hodarchaeota + eukaryotes), and from Alphaproteobacteria 2.58–2.12 Ga (Supplementary Fig. 16, Supplementary Data 10). Plastids diverged from free-living Cyanobacteria 2.14–1.73 Ga (Fig. 5A, C, Supplementary Fig. 16, Supplementary Data 10) and we inferred LECA to have originated 1.93–1.84Ga (Supplementary Fig. 16, Supplementary Data 10). These revised ages for key nodes in the species tree provide a timeline to study ATP synthase diversification in the context of cellular evolution: the split between the catalytic and non-catalytic ATP synthase subunits (4.52–4.46 Ga) likely predates (or at the latest was contemporary with) LUCA (4.52–4.32 Ga), while the divergence into F1- and A1/V1-types within the catalytic (4.52–4.38 Ga) and noncatalytic (4.52–4.42 Ga) clades overlaps in time with LUCA (4.52–4.32 Ga) and LBCA (4.49–4.05 Ga) but predates LACA (3.95–3.37 Ga) by more than 0.5 Gyr (Fig. 5B, Supplementary Figs. 18–19, Supplementary Data 10). An early origin of the A/V-type ATP synthases is a prerequisite for their presence in LBCA. If the split between F- and A/V-types corresponds to the speciation of Archaea and Bacteria, an older age for the A1/V1 clade compared to crown Archaea (LACA) might hint at a sampling or extinction “bottleneck” on the stem lineage leading to extant Archaea (Fig. 4B).

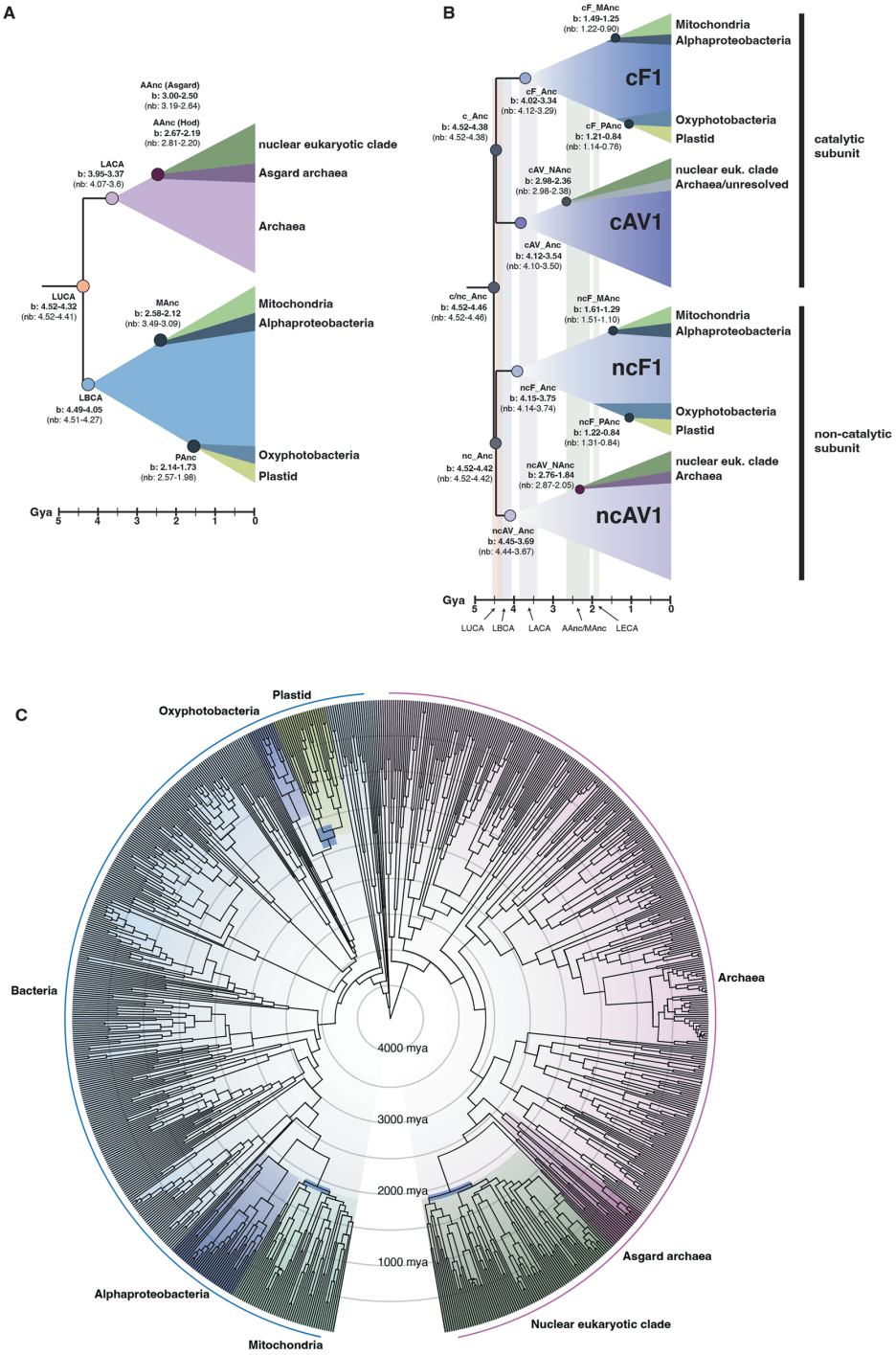


Fig. 5 | Timing of cellular evolution across the tree of life based on a cross-braced dated ribosomal species tree and ATP synthase gene tree. A Suggested timing of key evolutionary events based on a

schematic ribosomal species tree. **B** Suggested timing of key evolutionary events based on a schematic ATP synthase gene tree. **C** Dated cross-braced ribosomal species tree (Edited2, see Methods) including nuclear, mitochondrial, and plastid eukaryotic homologs. See Methods for inference of the maximum-likelihood concatenated ribosomal species phylogeny and constraints (Edited2). The alignment contained 863 sequences and was trimmed with TRIMAL *(129)* (alignment length = 2133 amino acids), and the maximum-likelihood phylogeny was inferred using IQ-TREE2 v2.1.2 the LG+C60+R+F model *(111, 122, 123)*. Abbreviations: b braced, nb non-braced, MAncL shared ancestor of mitochondria and closest alphaproteobacterial sister lineage, AAnc: shared ancestor of eukaryotic host lineage and closest asgardarchaeal sister lineage; Hod, Hodarchaeales; PAnc, shared ancestor of plastid and closest cyanobacterial sister lineage; c catalytic; nc non-catalytic; F F-type ATP-synthase; AV AV-type ATP-synthase.

DISCUSSION

The results of our analyses confirm that the A/V-type ATP synthase was present in LACA (*67*) and the F-type ATP synthase was in LBCA (*19–22*). They also revealed that A/V-type ATP synthases are broadly distributed in Bacteria and might have already been present in LBCA (Fig. 1, 3A, Supplementary Fig. 2, Supplementary Figs. 6–8, Supplementary Fig. 10). Previous analyses suggested that the acquisition of A/V-type ATP synthases in Bacteria occurred via HGT from hyperthermophilic Archaea (*33*, *35*, *53*), despite the observation that many mesophilic Bacteria also contain A/V-type ATP synthases (Figs. 1, 3A, Supplementary Fig. 2). In contrast, only three archaeal genomes (all within the genus *Methanosarcina*) appear to encode F-type ATP synthases (Figs. 1, 3A, Supplementary Fig. 2, Supplementary Figs. 4–5) belonging to a family of ATP synthases known as N-ATPases. The latter represent a distinct horizontally-acquired F-type ATP synthase which exists in addition to a bona fide F- or A/V-type ATP synthase in Bacteria and Archaea (*56*). Experimental studies revealed that the N-ATPase of *M. acetivorans* is not required for growth (*55*, *56*) and the function is debated.

ATP synthase evolution in Bacteria seems to be driven by frequent transfers from Archaea to Bacteria. Alternatively, A/V-type ATP synthases may already have been present in LBCA or LUCA and subsequently lost in many bacterial lineages. The latter possibility is in line with a scenario in which transfers from Bacteria to Archaea have been more common during evolution (*77*), but requires a loss of the F-type ATP synthase along the branch leading to Archaea (Fig. 4C). In agreement with this, the duplication giving rise to the catalytic and non-catalytic subunits (4.52–4.42 Ga 95% highest posterior density, Supplementary Figs. 18–19, Supplementary Data 10) and the divergence into F- and A/V-lineages within the catalytic and non-catalytic clades (4.52–4.38 Ga 95% highest posterior density, Supplementary Figs. 18–19, Supplementary Data 10) were inferred to have occurred very early in the history of cellular life, prior to, or overlapping with, LUCA (4.52–4.32 Ga, Fig. 5A, C, Supplementary Fig. 16, Supplementary Data 10). This may seem at odds with previous inferences (*10*, *29*, *32*, *78*) and suggestions of deep divergences between Archaea and Bacteria coinciding with distinct informational processing machinery (*19*), ATP synthases, and membrane lipids (*36*, *79*). However, the 'lipid divide' (*80*) appears less pronounced than assumed previously (*81*, *82*) and LUCA may have had the mevalonate pathway (*83*, *84*) and been able to synthesize bacterial and

archaeal-type lipids (*21*, *85*). The occurrence of the F- and/or A/V-type ATP synthases in modern Bacteria, suggests that their membrane lipids are compatible with either ATP synthase type. It is unclear whether the lack of Archaea with a complete substitution of an A/V- with F-type ATP synthase can be explained by constraints imposed by archaeal membrane lipid composition. Alternatively, it is possible that while the diversification of the catalytic and non-catalytic subunits into the respective F1 and A1/V1 families may have predated LUCA, the hexameric headpieces may have functioned independently from extant membrane components (*29*). In this scenario, the evolution of the membrane components could have occurred later, potentially in conjunction with the speciation of Bacteria and Archaea and the 'lipid divide' (*36*, *79*).

Despite the wide distribution of ATP synthases across cellular life, our analyses revealed that many DPANN Archaea and CPR Bacteria may have minimal complexes, as is the case in the DPANN *Nanoarchaeum equitans* (*86*), or even lack all genes for an ATP synthase complex (Fig. 1, Supplementary Fig. 2, Supplementary Data 4). This finding suggests that ATP synthases are not as essential as previously assumed (*32*, *33*, *45*, *65*) with loss in DPANN and CPR lineages likely being the result of genome streamlining processes consistent with their predicted host-dependent lifestyles (*58*, *59*). For instance, various members of DPANN lack several ATP synthase subunit homologs (60/103) while others encode homologs clustering with other DPANN or potential hosts (Fig. 1, Supplementary Figs. 6–10, Supplementary Data 4). We observed putative symbiont-host gene transfers between acidophilic Micrarchaeota and their hosts belonging to the Thermoplasmata (*87*) consistent with work supporting extensive HGT among ATP synthase genes of acidophilic archaeal lineages (*60*) (Supplementary Figs. 6–10, Supplementary Data 4). Furthermore, our trees indicate HGT and/or compositional attraction (*88*) of the *c*A1V1 subunit between the symbiotic Nanohaloarchaeota and their halobacterial hosts (Supplementary Fig. 6). The evolution of ATP synthase genes underpins debate over the phylogenetic placement of Nanohaloarchaeota (*45*), originally placed as the sister-group to Halobacteria (*45*, *89*) but later recovered as a member of DPANN (*16*, *17*, *90–92*). Recently, Feng and coworkers found that the catalytic and non-catalytic subunits of the A/V-type ATP synthase of Nanohaloarchaeota form sister-groups to halobacterial homologs (*45*). Although their concatenated species trees placed Nanohaloarchaeota with other DPANN (*45*), the authors argued that this placement was an artifact due to compositional biases in the concatenated dataset, with the ATP synthase gene tree recording the true organismal history. By contrast, Wang et al. (2019) (*60*) suggested that the incongruence of the species and ATP synthase gene trees for halophilic Archaea result from the HGT of an ATP synthase operon from Halobacteria into the common ancestor of Nanohaloarchaeota. Ecological association and symbiotic interactions between these organisms might have facilitated such a transfer. We include a larger representation of DPANN and metagenome-assembled genomes (GCA_003660905, GCA_003660865) belonging to a divergent sister lineage of the Nanohaloarchaeota (Fig. 1, Supplementary Fig. 2, 20), providing an opportunity to reconsider and distinguish hypotheses. Our results group Nanohaloarchaeota *nc*A1V1 subunits with Halobacteria, while in the *c*A1V1 subtree, Nanohaloarchaeota group with DPANN (including the sister lineage, Supplementary Figs. 6–10). Our analyses are most compatible with a

scenario in which the last common ancestor of Nanohaloarchaeal already possessed an ATP synthase complex inherited vertically from its DPANN relatives. The *nc*A1V1 subunit may have been replaced through HGT from a halobacterial host early during the lineage's evolution, potentially as an adaptation to halophily. Alternatively, it might be compositionally attracted to homologs of the Halobacteria as a result of convergent adaptations to halophily (*88*). This suggests that even genes whose synteny is conserved across lineages may be individually affected by HGT or evolutionary constraints. In such cases, the phylogenetic signal encoded in a larger number of marker genes may provide a more reliable estimate of the species tree.

Consistent with observations that niche expansion of Thaumarchaeota into acidic soils and high pressure oceanic zones was linked to their horizontal acquisition of a variant V-type ATP synthase operon (*60*), our results illustrate the potential role of symbiont-host gene exchange and environmental factors in ATP synthase evolution. Prospective studies focusing on genome evolution of DPANN archaea can help further assess the presence of ATP synthases and other metabolic components in the various DPANN ancestors and elucidate instances of transfer and loss of genes throughout DPANN diversification and adaptation to their respective symbiotic hosts.

Our molecular clock analyses suggest cross-bracing species nodes within gene trees is effective in propagating temporal information across the tree of life and improves the precision and accuracy of divergence time estimates for Archaea and Bacteria. Bracing resulted in higher estimated rates of molecular evolution overall (Supplementary Fig. 17, Supplementary Data 10), with the result that various deeper nodes of the tree were estimated to be slightly younger when compared with the un-braced analyses. This also includes LUCA, which has a mean age of 4.46 Ga in our cross-braced analyses and of 4.49 Ga in un-braced analyses. However, as observed in previous studies (*19*, *27*), the credibility interval associated with the LUCA still clashes against the root hard-maximum represented by the moon-forming impact even when implementing cross-bracing. This indicates that bracing helps to ameliorate, though not completely resolve, the problem of an under-calibrated clock inferring rates that are too low to account for the amount of genetic change that has occurred since the root of the universal tree (*19*, *93*). Some recent studies have reported moderately younger age estimates for LUCA: 4.05–3.42 Ga (*94*), or a range of values 4.48–3.93 Ga depending on conditions (*95*). An important driver of these differences is the choice of root maximum, which was younger in both studies (3.8–4.1 Ga (*95*) and 3.9 Ga (*94*)). In turn, also in those studies (*94*, *95*), the credibility interval for the age of LUCA clashes against the maximum used to calibrate the root node. This is consistent with previous work suggesting that the age of LUCA is sensitive to the root calibration used (*19*, *21*, *27*, *95*, *96*). We used the age of the Earth as our root maximum (the moon-forming impact at 4.52 Ga) because we are unaware of any compelling evidence for a younger maximum on the age of extant life (Supplementary Material). Thus, while the precise age of life and of LUCA remains uncertain, the inferred ages of LUCA and the early ATP synthase duplicates seem to imply a very high rate of evolutionary innovation during the earliest period of evolutionary history. Additional calibrations for deep nodes in the universal tree, along with

date estimates for other pre-LUCA paralogs, may help to dissect this key evolutionary period in higher resolution in future work (see Supplementary Discussion for further details about the resulting age estimates for major prokaryotic clades).

The LECA estimate (1.93–1.84 Ga, 95% highest posterior density) from our species tree analysis falls within published molecular clock estimates, placing LECA within a broad interval ~1–2.4 Ga (*25–27*, *97–99*) (Fig. 5A, C, Supplementary Fig. 16, Supplementary Data 10). More recent analyses have tended to resolve an older LECA, with ages closer to 1 Ga being less plausible on the basis of fossils from that period that can uncontroversially be assigned to crown Archaeplastida. These fossils include the green alga *Proterocladus antiquus* (1 Ga) (*100*) and the red alga *Bangiomorpha pubescens* (>1030 Mya) (*101*). The ages of some of these nodes, including LECA and particularly the last plastid common ancestor (LPCA), were inferred to be younger in the ATP synthase analysis (Fig. 5, Supplementary Figs. 16 and 18, Supplementary Data 10). In part, this may be due to the shorter alignment of ATP synthase (433-512AA, Supplementary Figs. 4–7, 10) and lower phylogenetic resolution (*102*) reducing the species tree calibrations and braces to the ATP synthase phylogeny through gene and species tree incongruence (Supplementary Information).

Our analyses are of interest for the timing of mitochondrial acquisition relative to other hallmark features of eukaryotes such as the nucleus (*103*, *104*) and help to explain the differences in the length of the stem between eukaryotic genes of archaeal and bacterial origin reported previously (*103*, *105*). Note that while the LECA nodes within the mitochondrial and nuclear lineages can be cross-braced to be contemporaneous, the lengths of the antecedent stems (i.e. the divergence times of the mitochondrial and nuclear lineages from their closest bacterial and archaeal relatives) might be very different (there should be no expectation that they are of equal antiquity). Our analyses support a moderately longer stem for the nuclear lineage (mean: 520.3 Ma, 291–789 Ma, 95% highest posterior density) than the mitochondrion (mean: 438.8 Ma, 233–682 Ma, 95% highest posterior density), suggesting the divergence of the nuclear lineage from the closest sampled Asgard archaea occurred before the divergence of the mitochondrial lineage from Alphaproteobacteria. However, the credible age ranges for these divergences overlap, therefore some additional factor (e.g. a faster evolutionary rate prior to LECA in eukaryotic genes of archaeal origin) may contribute to the observed differences in stem lengths (*103*, *105*). Interestingly, the inferred timescale is sensitive to the phylogenetic position of eukaryotes within Asgard archaea and Alphaproteobacteria: in an alternative analysis in which eukaryotes were placed sister to all Asgard archaea, and mitochondria within Alphaproteobacteria, the difference in stem group ages was more pronounced (mean 812.1 Ma, 95% highest posterior density 540–1105 Ma, nuclear stem: 310.1 Ma, mitochondrial stem: 150–508 Ma) (Supplementary Fig. 12). While this result tells us something about the shape of the tree of life it does not distinguish between hypotheses of an “early” or “late” mitochondrial acquisition. This is because these hypotheses make competing predictions about the order in which key eukaryotic features without direct correspondence to nodes in the tree were acquired relative to the mitochondrial endosymbiosis (Donoghue et al. 2023) (*106*).

CONCLUDING REMARKS

Our analyses provide insights into the diversification of the ATP synthase gene family and established age estimates for key nodes in the tree of life. Our results suggest that while LACA solely harbored an A1/V1-type ATP synthase, LBCA may already have encoded homologs of the head component of both the F- and a A/V-type ATP synthase. Studying how A/V-type ATP synthases function in Bacteria will help to explain the distribution we observed and the functional consequences of the ancient divergence between F- and A/V-type ATP synthases. In contrast to previous work, our inferences are consistent with the hypothesis that the divergence of the F1- and A/V1-type ATP synthase components may have predated LUCA. Furthermore, ATP synthase evolution supports scenarios on eukaryotic origins from an asgardarchaeal host (*3*, *4*, *13*, *14*) and alphaproteobacterial symbiont (*107*, *108*) and, together with our dated species tree, provide an updated timescale of cellular evolution, placing the origin of the eukaryotic cell into a geological context that can help to test eukaryogenesis models.

METHODS

SELECTION OF 800 TAXA COMPRISING THE BACKBONE GENOME REFERENCE DATASET

Archaeal reference genomes

A representative set of archaeal genomes was selected from a broad diversity of all archaeal genomes present in NCBI. A set of 51 marker proteins (*91*) was used to infer an initial concatenated phylogeny of 574 archaeal genomes meeting a threshold of >40% completeness and <13% contamination (Supplementary Data 1). Individual markers were aligned with MAFFT L-INS-i v7.407 (settings: –reorder) (*109*), trimmed using BMGE v1.12 (settings: -m BLOSUM30 -h 0.55) (*110*) and concatenated with a custom script (catfasta2phyml.pl; https://github.com/nylander/catfasta2phyml). A phylogenetic tree was generated with IQ-TREE v1.6.7 (settings: -m LG+C60+F+R -bb 1000 -alrt 1000) (*111*).

Based on this tree, 350 archaeal genomes were subsampled to evenly represent archaeal phylogenetic diversity (Supplementary Data 1). Type-strains were preferentially selected, while high quality metagenome assembled genome and single cell assembled genomes were selected based on completeness and contamination levels.

Bacterial reference genomes

The bacterial reference backbone, prioritizing type-strains and reference genomes, but also high-quality metagenome assembled genomes and a subselection of representatives from candidate phyla, was derived using an initial phylogeny of bacterial genomes available in NCBI as described above. Homologs of a conserved set of 29 marker proteins, i.e. a subset of 48 single-copy marker proteins previously defined in Zaremba-Niedzwiedzka et al. (2017) (*4*) were identified in those bacterial genomes, aligned using MAFFT v7.407 (settings: –reorder)

(*109*), trimmed using BMGE v1.12 (settings: -m BLOSUM30 -h 0.55) (*110*) and concatenated to reconstruct a phylogenetic tree using IQ-TREE v1.6.7 (settings: -m LG+G -bb 1000 -alrt 1000) (*111*). We subsampled the concatenated phylogeny for 349 bacterial genomes that represent known bacterial genomic diversity, ensuring selection of major bacterial taxonomic clades. The genome of *Schaalia odontolytica* ATCC 17982, which represents the host of members of the Saccharibacteria (formerly phylum TM7) (*112*, *113*), was downloaded from NCBI in 2020 and manually added to the bacterial backbone dataset (Supplementary Data 1).

Eukaryote reference genomes

A set of 100 published genome-wide datasets (genomes and, for lineages lacking complete genomes, largely complete transcriptomes) were sampled to represent the major lineages of eukaryotes (Supplementary Data 1, Supplementary Data 11). We also included sequences from the unpublished *Diplonema papillatum* genome project, with the permission of the sequencing consortium (see Acknowledgements).

Functional annotations

To identify sequences of ATP synthase subunits within all genomes in the 800-backbone set, all protein coding sequences were annotated using the KEGG and COG databases. Sequences were compared to KO profiles within the KEGG Automatic Annotation Server (KAAS, downloaded April 2019) (KAAS; downloaded April 2019) (*114*), to COG profiles within the NCBI COG database (downloaded May 2020) (*115–117*), and to Pfam profiles in the Pfam database (Release 34.0) (*118*). KOs and COGs were assigned using hmmsearch v3.1b2 (settings: –tblout sequence_ results.txt -o results_all.txt –domtblout domain_results.txt –notextw -E 1e-5) (*119*). Pfams were assigned using hmmsearch v3.1b2 (settings: --tblout sequence_results.txt -o results_all. txt --domtblout domain_results.txt --notextw -E 1e-10) (*119*).

INFERENCE OF A CONCATENATED SPECIES PHYLOGENY INCLUDING ARCHAEA, BACTERIA, AND EUKARYOTES

Marker gene homology search

A concatenated phylogeny of the 800 bacterial, archaeal, and eukaryotic genomes included in this study was inferred using a previously defined set of 27 single-copy marker genes (*19*) (Supplementary Fig. 20, Supplementary Data 12). To collect the corresponding homologs, the 800 reference genomes were queried against all COG HMM profiles with a custom script built on the hmmsearch [options] <reference genomes> <hmmfile> algorithm (*120*): hmmsearchTable Whole_ArcBacEuk_800_vs2_clean.faa NCBI_COGs_Oct2020.hmm -E 1e-5 > 1_Hmmersearch/HMMscan_Output_e5 (HMMER v3.3.2) (*121*), and all homologs corresponding to the 27 single-copy marker genes were identified, cleaned, and parsed. The approaches used to identify the appropriate homologs for prokaryotes and eukaryotes are described below.

Selection of prokaryotic homologs

For prokaryotes, the best-hit sequences were selected based on e-value and bitscore and the corresponding protein sequences were extracted from the reference genome backbone. Protein sequences assigned to each marker gene were aligned using MAFFT L-INS-i v7.453 (settings: --reorder) (*109*) and trimmed using BMGE v1.12 (settings: -t AA -m BLOSUM30 -h 0.55) (*110*). Maximum-likelihood phylogenies with ultrafast bootstrap approximation (UFBoot) for each single-copy marker gene were constructed using IQ-TREE2 v2.1.2 (settings: -m LG+G -wbtl -bb 1000 -bnni) (*111*, *122*, *123*). Individual marker gene trees were manually inspected for domain-level monophyly, the presence of paralogous protein families, and signs of contamination including LBA and horizontal gene transfer (HGT) (Supplementary Data 1, Zenodo data repository: https://doi.org/10.5281/zenodo.10012837 (*74*)). Marker genes, where domain-level lineages were paraphyletic were excluded and sequences with indications of LBA, HGT, and paralogy were manually removed using a custom script: remove_seq_with_specific_header3.py.

Selection of nuclear eukaryotic homologs

To distinguish between the nuclear, plastid, and mitochondrial homolog and select the correct eukaryotic representative sequence, we collected all eukaryotic hmmsearch hits and downsampled them with CD-HIT v4.7 using a threshold of 90% sequence identity (*124*, *125*). The filtered eukaryotic sequences were combined with the previously inspected prokaryotic sequences and all sequences for each single-copy marker gene were aligned using MAFFT L-INS-i v7.453 (settings: --reorder) (*109*), and trimmed using BMGE v1.12 (settings: -t AA -m BLOSUM30 -h 0.55) (*110*). Single gene phylogenies were inferred using FastTree (settings: -lg) (*126*). KEGG and Pfam annotations (see above) were mapped to the tips of the eukaryotic sequences for manual inspection of multiple paralogs per taxon. First, the eukaryotic sequences were inspected by removing any sequence failing monophyly (i.e., HGTs in prokaryotic clades) or not clearly derived from the nuclear source (i.e., the plastid and/or mitochondrial sequences). Duplicate nuclear eukaryotic sequences were filtered in a two-step procedure: (1) if duplicate sequences are monophyletic, select a single representative based on protein annotation consistent with the identity of the single-copy marker gene, and (2) if duplicate sequences are paraphyletic, remove taxon completely from the single-copy marker gene. Any representatives with fewer than 65% of the marker genes (20 taxa removed, 80 eukaryotes in total) were removed from this analysis (Supplementary Data 13).

Inspection of final marker gene sequence sets

The final set of eukaryotic nuclear sequences were combined with the previously inspected sequences for Archaea and Bacteria (see above) and aligned using MAFFT L-INS-i v7.453 (*109*) 109, trimmed with BMGE v1.12 (settings: -t AA -m BLOSUM30 -h 0.55) (*110*) 110, and single gene trees were inferred using maximum-likelihood with UFBoot approximation methods in IQ-TREE2 v2.1.2 (settings: -m LG+G -wbtl -bb 1000 -bnni) (*111*, *122*, *123*). Upon inspection of single gene trees including homologs from Archaea, Bacteria and eukaryotes, six single-copy markers (COG0064, COG0085, COG0086, COG0202, COG0480, and COG5257 (Supplementary

Data 12) were flagged for removal (e.g. lack of clear nuclear paralog, or absence of archaeal or bacterial sequences in the tree).

Inference of the concatenated phylogeny

Alignments for the 21 single-copy marker genes were generated and trimmed following the approaches outlined above and individual marker alignments were concatenated using the script catfasta2phyml.pl (https://github.com/nylander/catfasta2phyml). The final concatenated sequence alignment contained 3367 positions and was used to infer maximum-likelihood phylogenies using varying models of evolution in IQ-TREE2 v2.1.2 (settings: -m LG+C60+R+F *or* LG+C20+R+F -bb 1000 -alrt 1000) (*111*, *122*, *123*). We examined the statistical support for topologies of the two concatenated species trees inferred under different models of evolution (LG+C60+R+F and LG+C20+R+F, see above, Supplementary Data 14, Zenodo data repository: https://doi.org/10.5281/zenodo.10012837 (*74*)) using the approximately unbiased (AU) test implemented in IQ-TREE2 v2.1.2 (settings: -s 21eLife_ArcBacEuk_wHuber_vs1.faa -m [LG+C20+R+F/LG+C60+R+F] -z 21eLife_ArcBacEuk_wHuber_vs1_bothtrees.treefile -pre [C20/C60] -n 0 -zb 10000 -au (*111*, *123*, *127*). Results are shown in Supplementary Data 15. Despite statistical exclusion of the LG+C20+R+F topology, we chose to use this tree for phylogenetic interpretation because the placement of key lineages such as the Asgard archaea and CPR, is most consistent with recent evidence (*4*, *19*, *21*, *22*) (Supplementary Fig. 20, Supplementary Data 15).

CONSTRUCTING A RIBOSOMAL MARKER PHYLOGENY INCLUDING NUCLEAR, MITOCHONDRIA, AND PLASTID HOMOLOGS

Selection of eukaryotic nuclear, mitochondrial, and plastid homologs

Eukaryotes encode two or more ribosomes of distinct prokaryotic origins, i.e., archaeal, alphaproteobacterial and, in the case of the presence of a plastid, a cyanobacterial origin (i.e. the nuclear, mitochondrial, and plastid, respectively). A concatenated phylogeny including, if identified, the nuclear, mitochondrial, and plastid ribosomal protein homologs for each eukaryote, was inferred for molecular dating and bracing analyses. To this end, we constructed single gene trees of the 15 ribosomal marker genes (subset of the 21 single-copy marker genes described above) which included Archaea, Bacteria, and all eukaryotic homologs (i.e., the nuclear, mitochondrial, and plastid). Note that the nuclear eukaryotic sequences were the same set of sequences reported in the final inspection of the concatenated species phylogeny (see above). To identify the plastid homologs, we selected the monophyletic clade of eukaryotic sequences affiliated with the Cyanobacteria. The mitochondrial sequences appeared to demonstrate variable placements with some affiliating with the alphaproteobacteria and others branching basally in the Bacteria. Therefore, we made our sequence selection based on the position of known mitochondrial genes of the type-species *Homo sapiens* and *Saccharomyces cerevisiae*. First, we manually located *H. sapiens* in the phylogenies and searched the protein accession in Uniprot and/or NCBI (*128*) to confirm sequence annotation and identity as a mitochondrial sequence. In the absence of a *H. sapiens* homolog, we used *S. cerevisiae* mitochondrial homologs. Of the 15 ribosomal

markers, three had no distinguishable mitochondrial homolog for either type-species and were dropped from the dataset, resulting in 12 ribosomal markers (Supplementary Data 16). All eukaryotic sequences, that clustered with the *H. sapiens/S. cerevisiae* homolog and grouped with alphaproteobacteria or basally in the phylogeny, were selected for subsequent analyses. Selected sequences were de-replicated using the following criteria: (1) if paralogous sequences are monophyletic retain one homolog based on annotation or manual selection, and (2) if paralogous sequences are paraphyletic remove all sequences from that organism. Dereplicate sequences marked for removal are in Supplementary Data 16. Gene trees with selected sequences have been deposited in our data repository at Zenodo: https://doi.org/10.5281/zenodo.10012837 (*74*).

Ribosomal protein homologs were then annotated based on their distinct origin (nuclear, mitochondrial, plastid) and the percent distribution of homologs across the 12 ribosomal markers by eukaryotic taxon was calculated. Only taxa that had at least 50% of the markers of nuclear, mitochondrial, or plastid origin were retained, resulting in 88 nuclear taxa, 50 mitochondrial taxa, and 25 plastid taxa (Supplementary Data 13). The eukaryotic GenomeIDs for each sequence were annotated with the suffix of the origin (i.e., EukGenome_nuclear, EukGenome_mito, etc.) for downstream concatenation. In total, the sequence sets for the 12 ribosomal markers contained archaeal and bacterial homologs, and the eukaryotic nuclear, mitochondrial, and plastid sequences, respectively. Alignments were generated using MAFFT L-INS-i v7.453 (settings: --reorder) (*109*) and trimmed with TRIMAL v1.2rev59 (settings: -gappyout) (*129*). The alignments of the 12 ribosomal markers were concatenated using the script catfasta2phyml.pl (https://github.com/nylander/catfasta2phyml) and the final concatenated alignment contained 2133 sites.

Inference of concatenated phylogenies

A maximum-likelihood phylogeny was inferred using IQ-TREE2 v2.1.2 (settings: -m LG+C60+R+F -bb 1000 -alrt 1000) (*111*, *122*, *123*) (Supplementary Fig. 21).

ASSESSING DISTRIBUTION OF ATP SYNTHASE GENES ACROSS 800 TAXA BACKBONE

We performed a comparative genomic analysis of the distribution of ATP synthase genes across the 800 taxa included in this study. COG families corresponding to each subunit of the ATP synthase (Supplementary Data 3) were extracted from the 800 reference genomes. Results were compiled, counted in R v4.1.1 (Supplementary Data 4). The count table was converted to a binary presence/absence matrix that was summarized using the ddply function of the plyr package (v1.8.6) by the respective phylogenetic clustering methods: (1) species-level according to order of individual species in the inferred concatenated phylogeny (BinID and Tip_Order, Supplementary Data 17), and (2) class- and phylum-level for Archaea and phylum-level for Bacteria corresponding to clade clustering in the concatenated phylogeny (CladeCluster and Clade_Order, Supplementary Data 17). The percentage distribution of subunits within each phylogenetic cluster was visualized in a bubble plot implemented using the ggplot function

with geom_tile and facet_grid from the ggplot2 package v3.3.5. The binary presence/absence of subunits by species was visualized with the ggplot function using geom_point and facet_grid from ggplot2 v3.3.5. All heatmaps and bubble plots were manually merged with the corresponding concatenated species tree in Adobe Illustrator CC v22.0.1.

Curation of eukaryotic hits

We conducted an additional step of quality control to curate eukaryotic protein sequences potentially corresponding to the key ATP synthase subunits highlighted in Supplementary Data 3. All eukaryotic proteins suggested to represent homologs of ATP synthase COGs (Supplementary Data 3) were identified in the protein annotation table (Supplementary Data 2) and the corresponding sequences were queried against the NCBI non-redundant (NCBI_nr) database using diamond blast v2.0.8 (settings: diamond blastp -q ${sample}_seqs.faa --more-sensitive --evalue 1e-5 --threads 20 --seq 100 --no-self-hits --db nr.dmnd --taxonmap prot.accession2taxid.gz --outfmt 6 qseqid qtitle qlen sseqid salltitles slen qstart qend sstart send evalue bitscore length pident staxids (*130*). We ranked the hits by e-value and bitscore and collected the (up to) top 10 hits per accession. Taxonomic information was mapped to the table using the NCBI taxonomy corresponding to the taxid. Domain identity for the top 10 hits per protein sequence were summarized and any sequence with ≥50% hits to Bacteria was considered putative bacterial contamination and flagged for removal (Supplementary Data 4). In total, 326 accessions were removed and not considered for the presence-absence analysis of the ATP synthase subunits (Supplementary Data 4). Putative contamination was also inspected in the protein sequences used to infer ATP synthase gene phylogenies and four sequences have been highlighted as putative bacterial contamination (Supplementary Data 4, Supplementary Figs. 4–5, Zenodo data repository: https://doi.org/10.5281/zenodo.10012837 (*74*)).

PHYLOGENETICS OF ATP SYNTHASE SUBUNITS

Sequence retrieval and selection

Interpro domains that characterize the protein families corresponding to the subunits present in the catalytic (R1) domain of the F-Type and A/V-Type ATP synthases were selected at the family-level (*131*) and include: ipr005294 (F-Type alpha, hereafter *nc*F1), ipr005722 (F-Type beta, hereafter *c*F1), ipr022878 (A/V-Type A, hereafter *c*A1/V1), and ipr022879 (A/V-type B, hereafter *nc*A1/V1). All protein sequences assigned to the corresponding interpro domains were extracted from the UniProt Knowledge Base (*128*), and were searched against the 800 reference genomes using DIAMOND v0.9.22.123 (settings: blastp -p 4 -f 6 qseqid stitle pident length mismatch gapopen qstart qend sstart send e-value bitscore) (*130*). Top hits were selected based on best e-value and sequence identity, and all unique protein accessions (from the 800 reference taxa) were used to extract the amino acid sequences from the 800-genome reference dataset. Sequences with undefined characters (i.e., X, x) and/or outside of the average sequence length of homologs, i.e., 300–675 bp, were filtered from the sequence sets. To avoid highly similar duplicates, sequences with 99–100% identity were removed using CD-HIT (*124*, *125*). Additionally, for consistency with the concatenated species phylogeny (see

above), eukaryotic taxa that fell below the 65% marker gene presence cutoff (20 eukaryotic taxa, Supplementary Data 13) were removed from the single-subunit sequence sets.

ATP SYNTHASE SUBUNIT PHYLOGENIES: *c*F1, *nc*F1, *c*A1V1, *nc*A1V1

A series of seven different sequence sets were generated for analysis:

1. Single subunits sets F1-alpha (*nc*F1), F1-beta (*c*F1), A1/V1A (*c*A1V1), and A1/V1B (*nc*A1V1) (four in total)
2. Combined orthologous subunits for outgroup rooting: F1A+A1/V1B (*nc*F1+*nc*A1V) and F1B+A1/V1A (*c*F1+*c*A1V1)
3. All four subunits combined

Potential duplicates were removed from the combined sets using CD-HIT v4.7 with a 100% identity (settings: cd-hit -1) (*124*, *125*), sequences were aligned using MAFFT L-INS-i v7.453 (settings: --reorder) (*109*) and trimmed using BMGE v1.12 (settings: -m BLOSUM30 -h 0.55) (*110*). The best-fit model was determined using the Model Finder Plus tool implemented in IQ-TREE2 v.2.1.2 (settings: -m MFP -mset LG -madd LG+C10,LG+C20,LG+C30,LG+C40,LG+C50,LG+C60,LG+C10+R+F,LG+C20+R+F,LG+C30+R+F,LG+C40+R+F,LG+C50+R+F,LG+C60+R+F --score-diff all -bb 1000 -alrt 1000 -bnni -wbtl) (*111*, *122*, *123*, *132*) and the best-fitting model for each gene tree was selected based on the Bayesian Information Criterion (BIC, Supplementary Data 8) and used to infer the maximum-likelihood phylogeny. Genome identifiers containing the GenomeID and protein accession were converted to a modified NCBI taxonomic string using an in-house script (Replace_tree_names.pl, https://github.com/ndombrowski/Phylogeny_tutorial/tree/main/Input_files/5_required_scripts). Trees were viewed in FigTree v1.4.4, and inspected for topological congruence and phylogenetic artifacts to iteratively improve the sequence selection, i.e. to exclude distant paralogs and sequences subject to long branch attraction (LBA) (*133*).

TRACING THE PHYLOGENETIC RELATIONSHIPS OF EUKARYOTIC F-TYPE ATP SYNTHASES

To better resolve the evolutionary origins of eukaryotic F-type ATP synthases we constructed phylogenies with a subset of sequences which included eukaryotic sister lineages (alphaproteobacteria and cyanobacteria for the mitochondrial and plastid-type F-ATP synthase, respectively) as well as an outgroup lineage (hereafter: plastid and mitochondrial subsets). For the plastid origin dataset, we selected all eukaryotic ATP synthase sequences from the *nc*F1 and *c*F1 subunit gene trees (see above) that clustered with the Cyanobacteria and added cyanobacterial and melainabacterial homologs. Similarly, the mitochondria origin subset was generated by collecting all eukaryotic ATP synthases sequences from the *nc*F1 and *c*F1 subunit gene trees that clustered with alphaproteobacterial homologs and adding additional alphaproteobacterial sequences and gammaproteobacterial homologs. Note that for both the plastid and mitochondrial sets, we used an expanded selection of cyanobacteria and alphaproteobacteria, respectively. Sequence selections were filtered to retain high quality sequences without ambiguous amino acids (i.e., X and x, etc.) and within the range of 450–550 bp. Closely related paralogous sequences were removed using CD-HIT v4.7 (set-

tings: -c 0.99) (*124*, *125*) and alignments were generated using MAFFT L-INS-i v7.453 (*109*) and trimmed using BMGE (settings: -m BLOSUM30 -h 0.55) (*110*). We inferred phylogenies using the best-fit model determined in the Model Finder Plus tool in IQ-TREE2 v2.1.2 (settings: -m TESTONLY -mset LG -madd LG+C10,LG+C20,LG+C30,LG+C40,LG+C50,LG+C60,LG+C10+R+F,LG+C20+R+F,LG+C30+R+F,LG+C40+R+F,LG+C50+R+F,LG+C60+R+F --score-diff all) (*111*, *122*, *123*, *132*) and the maximum-likelihood trees were constructed in IQ-TREE v1.6.10 using the best-fit model based on the BIC (*111*) (Supplementary Data 8).

Additionally, we used Bayesian analysis to further verify the placement of eukaryotic F1-type ATP synthase sequences amongst the proposed sister lineages. Due to computational limitation, we downsampled the taxa subsets containing eukaryotes, alphaproteobacteria, and gammaproteobacteria to a maximum of 250 taxa (*nc*F1: 211, *c*F1:185 sequences). Sequences were cleaned, filtered, de-replicated, aligned, and trimmed using the same conditions described above. Bayesian phylogenies were constructed using PhyloBayes-MPI (version 1.5) using the CAT-GTR model with four discrete gamma categories for rates across sites; for each alignment, four independent Markov Chain Monte Carlo (MCMC) chains were run. Each chain was run over 100,000 iterations (or until convergence). Convergence was evaluated using the bpcomp and tracecomp tools within PhyloBayes-MPI, with 1000 generations discarded as burn-in and sub-sampling every 10 trees. The final consensus trees were generated through bpcomp using the same settings.

ANCESTRAL SEQUENCE RECONSTRUCTION

Sequence alignments and the accompanying maximum-likelihood trees for the ATP synthase subunits, the orthologous pairs, and the set of four combined subunits were used to reconstruct the ancestral protein sequences. For ancestral sequence reconstruction we used a tool implemented in IQ-TREE2 v2.1.2 (settings: -m [model] -asr -te [maximum likelihood tree] -keep_empty_seq) (*123*). Ancestral sequences were determined based on the proposed amino acid states at specified node positions in the rooted combined ATP synthase protein tree (Supplementary Fig. 10, Supplementary Data 7).

CONSERVED NUCLEOTIDE-BINDING MOTIFS

Untrimmed and trimmed alignments of F- and A/V-ATP synthase subunits from the 800 reference genomes (see above) were manually inspected in Jalview v2.10.5 (*134*) for the presence of the WalkerA (P-loop) motif (*43*). The signature nucleotide-binding motif is characterized by the amino acid sequence: *GXXXXGK(T/S)* where X denotes any amino acid. The WalkerA motif sequence segment was extracted from the full alignment and used to generate conserved motif logos in WebLogo3 v3.7.4 (*135*) (http://weblogo.threeplusone.com/).

DATING THE TREE OF LIFE AND ATP SYNTHASE PHYLOGENIES

The absolute time calibrations used in the dating analysis are detailed in the Supplementary Discussion. As the fossil evidence with which to constrain early microbial evolution is limited, we also used cross-bracing (*24*) to propagate the available calibrations across the tree,

implemented in McmcDate (see below). In particular, we braced the LECA node that appears in the nuclear and mitochondrial clades (setting their ages to be the same), along with all calibrated nodes within eukaryotes that were present in two or more of the eukaryotic clades (that is, we braced all nodes within eukaryotes where a geological calibration was applied). Finally, we implemented a relative constraint (*49*) that the crown plastids must be younger than archaeal- and mitochondrial LECA (Supplementary Discussion).

We used McmcDate (https://github.com/dschrempf/mcmc-date) for molecular dating. McmcDate approximates the phylogenetic likelihood using a multivariate normal distribution obtained from an estimate of the posterior distribution of trees with branch lengths measured in average number of substitutions per site. We estimated the posterior distribution of trees in a previous step. For this previous step we used PhyloBayes (LG+G4 model) and a fixed phylogeny, as described above. We sampled 10,000 values of the posterior distribution of trees and observed good convergence with estimated sample size (ESS) values of around 8000.

Using McmcDate, we sampled 12,000 time trees. We used a birth-death tree prior on the time tree, and an uncorrelated log normal relaxed molecular clock model. We calibrated node ages using uniform distributions with bisected normal distributions at the boundaries. Similarly, we constrained the node order using the tails of normal distributions. We set the steepness of the boundaries individually depending on the quality and certainty of the auxiliary data. In a similar way, we used normal distributions to brace nodes. ESS values indicated good convergence and ranged from 3000 to 6000.

Application of fossil calibrations to the inferred maximum-likelihood ribosomal species tree (see above, Supplementary Fig. 20) was limited due to poor resolution of within-Eukaryote relationships. To apply an extended set of fossil calibrations we fixed the within-eukaryote topology to reflect established relationships among supergroups (*136*) and to allow the within-eukaryote fossil calibrations to be applied to the tree (see calibrations justification, Supplementary Discussion, Supplementary Data 18–20). In addition to these eukaryotic constraints, one topology (hereafter, Edited1) placed the nuclear eukaryotic homologs as the sister lineage to all asgardarchaeal and the mitochondrial homologs as the sister lineage to a single *Neorickettsia* (Supplementary Fig. 12). We used a more conservative approach to investigate the timing of LECA via the nuclear and mitochondrial eukaryotic nodes by adding additional constraints (hereafter, Edited2) to position the nuclear homologs as the sister lineage to the Hodearchaea (formerly Heimdallarchaeota LC3), their predicted closest asgardarchaeal relatives (*4*, *137*), and the mitochondrial homologs sister to all alphaproteobacteria, consistent with previous work (*107*, *108*) (Supplementary Fig. 13). The focal analysis described here is derived from dating the Edited2 topology (Fig. 5C, Supplementary Figs. 13 and 16).

An Approximately-Unbiased (AU) test was applied to assess the statistical support for the different topologies inferred from the ribosomal species trees utilized for cross-calibrated dating. The AU test was implemented in IQ-TREE2 v.2.1.2 (*123*, *127*) (settings: iqtree2 -s

12Ribosomal_eLife_ArcBacEuk_gappyout_v1b.faa -m LG+C60+R+F -z Ribo_C60_trees.alltrees.treefile -n 0 -zb 10000 -au). Results are shown in Supplementary Data 21.

In formulating the calibrations (Supplementary Discussion), we followed the best practice principles set out in Parham et al. (2012) (*138*). However, these were designed with animal and plant fossil-based calibrations and not all of the principles are applicable to calibrations of microbial clades which often lack phenotypic synapomorphies, let alone diagnostic characters that are preserved in fossil remains. Furthermore, the calibrations for many clades rely on geochemical evidence of microbial metabolisms, manifest as isotope fractionation or oxidation states of redox sensitive mineral species. Consequently, we have adapted the best practice principles to suit the nature of the calibrations. Novel calibrations are justified in full; we indicate the source of calibrations that are justified elsewhere, providing notes where they have been adapted for different clades or where the dating has changed with the revision of the geologic timescale (Supplementary Discussion).

GENE TREE–SPECIES TREE RECONCILIATION USING AMALGAMATED LIKELIHOOD ESTIMATION (ALE)

Ultrafast bootstraps (UFBoot) were inferred for each of the ATP synthase gene trees (see above) in IQ-TREE2 v2.1.2 (*111*, *122*, *123*), and the inferred maximum-likelihood concatenated species trees (see above). ALEobserve was used to convert bootstrap distributions into ALE objects, which were reconciled using ALEml_undated against each of the four species trees: those with eukaryotes, using the LG+C20+R+F and LG+C60+R+F model, and those lacking eukaryotic sequences, with the same two models (Supplementary Data 6). These four species tree topologies were also rooted in two different ways: a root between Archaea and Bacteria, and a root between Gracilicutes and all other taxa. Two approaches were taken using ALE for gene tree-species tree reconciliation. First, we used the default ALE parameters, i.e. inferring the probability that each subunit originated at the LUCA, LBCA and LACA nodes on the prior assumption that origination at any internal node of the species tree was equally likely. We also tested an alternative approach (*21*) in which the origination probability at the root (O_R) is different to the origination probability for all other internal nodes of the tree, with O_R estimated by maximum-likelihood. Reconciliation analyses were performed using ALE v1.0 (https://github.com/ssolo/ALE).

To compare support for the traditional Archaea-Bacteria root for the tree of life, and an alternative root within the Bacteria, we used gene tree-species reconciliation. We performed gene tree-species tree reconciliation using the species tree as described above as well as individual gene family subunit trees of ATP synthase: *nc*F1 (F1 alpha), *c*F1 (F1 beta), *c*A1V1 (A1/V1 A), and *nc*A1V1 (A1/V1 B), as well as three combined gene families, *nc*F1+*nc*A1V1, *c*F1+*c*A1V1, and all four families combined (Supplementary Data 6). Two taxon samplings were used as described above, one with 350 Archaea and 350 Bacteria only, and another with 350 Archaea, 350 Bacteria, and 100 eukaryotes. The summed gene family likelihoods of each ATP synthase subunit were compared using an AU test (*127*) as implemented in CONSEL (*139*) under a range

of conditions: species trees inferred under the LG+C20+R+F and LG+C60+R+F models; samples including and excluding eukaryotes, and two different root positions, one with the traditional root between Archaea and Bacteria, and the second with a within-Bacteria root on the branch leading to Gracilicutes.

DATA AVAILABILITY

All genomic data of Archaea and Bacteria analyzed are available at NCBI (Supplementary Data 1), while all eukaryotic genomic/transcriptomic material is deposited in our data repository at Zenodo (https://doi.org/10.5281/zenodo.10012837). Data generated in this study including single gene tree analyses and concatenated phylogenies (i.e., sequence files, alignments, and treefiles) have also been deposited in our data repository at Zenodo (https://doi.org/10.5281/zenodo.10012837) under the following license CC BY 4.0. Public databases are available as follows: ATP synthase Interpro domains were downloaded from Uniprot Knowledge Base (2019) (https://www.uniprot.org/), KO profiles downloaded from the KEGG Automatic Annotation Server in 2019 (https://www.genome.jp/tools/kofamkoala/), and the NCBI COG Database downloaded May 2020 (https://ftp.ncbi.nih.gov/pub/COG/COG2020/data/).

CODE AVAILABILITY

Workflows for annotations and phylogenies and custom R scripts to analyze and parse annotation data for figure generation have been deposited in our data repository at Zenodo (https://doi.org/10.5281/zenodo.10012837). We used the following published codes: Replace_tree_names.pl (https://github.com/ndombrowski/Phylogeny_tutorial/tree/main/Input_files/5_required_scripts), Mcmcdate (https://github.com/dschrempf/mcmc-date), catfasta2phyml.pl (https://github.com/nylander/catfasta2phyml).

ACKNOWLEDGEMENTS

This work was supported by the Simons Foundation (735929LPI, to A.S., https://doi.org/10.46714/735929LPI) and the Gordon and Betty Moore Foundation (GBMF9741 to T.A.W., A.S., G.J.S., D.P. and P.C.J.D) and the Gordon and Betty Moore Foundation's Symbiosis in Aquatic Systems Initiative (GBMF9346, to A.S.). Furthermore, this project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 947317, ASymbEL to A.S. and grant agreement No. 714774, GENECLOCKS to G.J.S.). Further, this work was supported by a Royal Society University Research Fellowship to T.A.W and the John Templeton Foundation (62220, to P.C.J.D., D.P. and T.A.W). Please note that the opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foun-

dation. We are also thankful for financial support from the Swedish Research Council (VR starting grant 2016-03559 to A.S.), the NWO-I foundation of the Netherlands Organisation for Scientific Research (WISE fellowship to A.S.). Finally, we are thankful for financial support from the Leverhulme Trust (RF-2022-167, to P.C.J.D.), the Biotechnology and Biological Sciences Research Council (BB/T012773/1, to P.C.J.D.) and the University of Bristol for a University Research Fellowship (to D.P.). We thank Gertraud Burger, Julius Lukes, Takeshi Nara, and other members of the Diplonema papillatum sequencing consortium for sharing data. We also want to thank Courtney Stairs, Andrew Roger, and Georg Hochberg for helpful discussions and/or feedback regarding eukaryotic metabolism and ancestral sequence reconstructions, respectively.

AUTHOR CONTRIBUTIONS

A.S. and T.A.M. conceptualized the study. T.A.M., E.R.R.M., T.A.W., D.S., L.L.S., N.D., G.J.S., A.A.D., and A.S. performed analyses and interpreted data. All authors contributed methods. D.P. and P.C.J.D. contributed data. A.S., T.A.W., G.J.S. and P.C.J.D. acquired funding. T.A.M. wrote the first draft with the help of A.S. and T.A.W. T.A.M., A.S., T.A.W. and E.R.R.M. together wrote the final draft. G.J.S., D.S., N.D., P.C.J.D. and D.P. contributed to the writing of the final manuscript and all authors read and approved the final version.

Competing interests
The authors declare no competing interests.

ADDITIONAL INFORMATION

SUPPLEMENTARY INFORMATION

The online version contains supplementary material available at https://doi.org/10.1038/s41467-023-42924-w.

All supplementary information and files be accessed here:
https://www.nature.com/articles/s41467-023-42924-w#Sec35

REFERENCES

1. A. Spang, T. A. Mahendrarajah, P. Offre, C. W. Stairs, Evolving Perspective on the Origin and Diversification of Cellular Life and the Virosphere. *Genome Biol. Evol.* **14** (2022).

2. L. A. Hug, B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D. A. Relman, K. M. Finstad, R. Amundson, B. C. Thomas, J. F. Banfield, A new view of the tree of life. *Nat Microbiol* **1**, 16048 (2016).

3. A. Spang, J. H. Saw, S. L. Jørgensen, K. Zaremba-Niedzwiedzka, J. Martijn, A. E. Lind, R. van Eijk, C. Schleper, L. Guy, T. J. G. Ettema, Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).

4. K. Zaremba-Niedzwiedzka, E. F. Caceres, J. H. Saw, D. Bäckström, L. Juzokaite, E. Vancaester, K. W. Seitz, K. Anantharaman, P. Starnawski, K. U. Kjeldsen, M. B. Stott, T. Nunoura, J. F. Banfield, A. Schramm, B. J. Baker, A. Spang, T. J. G. Ettema, Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).

5. A. Spang, L. Eme, J. H. Saw, E. F. Caceres, K. Zaremba-Niedzwiedzka, J. Lombard, L. Guy, T. J. G. Ettema, Asgard archaea are the closest prokaryotic relatives of eukaryotes, *PLoS genetics*. **14** (2018)p. e1007080.

6. C. Rinke, M. Chuvochina, A. J. Mussig, P.-A. Chaumeil, A. A. Davín, D. W. Waite, W. B. Whitman, D. H. Parks, P. Hugenholtz, A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nat Microbiol* **6**, 946–959 (2021).

7. T. A. Williams, C. J. Cox, P. G. Foster, G. J. Szöllősi, T. M. Embley, Phylogenomics provides robust support for a two-domains tree of life. *Nat Ecol Evol* **4**, 138–147 (2020).

8. Y. Liu, K. S. Makarova, W.-C. Huang, Y. I. Wolf, A. N. Nikolskaya, X. Zhang, M. Cai, C.-J. Zhang, W. Xu, Z. Luo, L. Cheng, E. V. Koonin, M. Li, Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature* **593**, 553–557 (2021).

9. L. Guy, J. H. Saw, T. J. G. Ettema, The archaeal legacy of eukaryotes: a phylogenomic perspective. *Cold Spring Harb. Perspect. Biol.* **6**, a016022 (2014).

10. W. F. Martin, S. Garg, V. Zimorski, Endosymbiotic theories for eukaryote origin. (2015). https://doi.org/10.1098/rstb.2014.0330.

11. E. V. Koonin, Origin of eukaryotes from within archaea, archaeal eukaryome and bursts of gene gain: eukaryogenesis just made easier? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140333 (2015).

12. P. López-García, L. Eme, D. Moreira, Symbiosis in eukaryotic evolution. (2017). https://doi.org/10.1016/j.jtbi.2017.02.031.

13. A. Spang, C. W. Stairs, N. Dombrowski, L. Eme, J. Lombard, E. F. Caceres, C. Greening, B. J. Baker, T. J. G. Ettema, Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. *Nat Microbiol* **4**, 1138–1148 (2019).

14. H. Imachi, M. K. Nobu, N. Nakahara, Y. Morono, M. Ogawara, Y. Takaki, Y. Takano, K. Uematsu, T. Ikuta, M. Ito, Y. Matsui, M. Miyazaki, K. Murata, Y. Saito, S. Sakai, C. Song, E. Tasumi, Y. Yamanaka, T. Yamaguchi, Y. Kamagata, H. Tamaki, K. Takai, Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**, 519–525 (2020).

15. P. López-García, D. Moreira, The Syntrophy hypothesis for the origin of eukaryotes revisited. *Nat Microbiol* **5**, 655–667 (2020).

16. C. Rinke, P. Schwientek, A. Sczyrba, N. N. Ivanova, I. J. Anderson, J.-F. Cheng, A. Darling, S. Malfatti, B. K. Swan, E. A. Gies, J. A. Dodsworth, B. P. Hedlund, G. Tsiamis, S. M. Sievert, W.-T. Liu, J. A. Eisen, S. J. Hallam, N. C. Kyrpides, R. Stepanauskas, E. M. Rubin, P. Hugenholtz, T. Woyke, Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).

17. C. J. Castelle, K. C. Wrighton, B. C. Thomas, L. A. Hug, C. T. Brown, M. J. Wilkins, K. R. Frischkorn, S. G. Tringe, A. Singh, L. M. Markillie, R. C. Taylor, K. H. Williams, J. F. Banfield, Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr. Biol.* **25**, 690–701 (2015).

18. C. T. Brown, L. A. Hug, B. C. Thomas, I. Sharon, C. J. Castelle, A. Singh, M. J. Wilkins, K. C. Wrighton, K. H. Williams, J. F. Banfield, Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).

19. E. R. R. Moody, T. A. Mahendrarajah, N. Dombrowski, J. W. Clark, C. Petitjean, P. Offre, G. J. Szöllősi, A. Spang, T. A. Williams, An estimate of the deepest branches of the tree of life from ancient vertically evolving genes. *Elife* **11** (2022).

20. N. Taib, D. Megrian, J. Witwinowski, P. Adam, D. Poppleton, G. Borrel, C. Beloin, S. Gribaldo, Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. *Nat Ecol Evol* **4**, 1661–1672 (2020).

21. G. A. Coleman, A. A. Davín, T. A. Mahendrarajah, L. L. Szánthó, A. Spang, P. Hugenholtz, G. J. Szöllősi, T. A. Williams, A rooted phylogeny resolves early bacterial evolution. *Science* **372** (2021).

22. C. A. Martinez-Gutierrez, F. O. Aylward, Phylogenetic Signal, Congruence, and Uncertainty across Bacteria and Archaea. *Mol. Biol. Evol.* **38**, 5514–5527 (2021).

23. P. Kapli, T. Flouri, M. J. Telford, Systematic errors in phylogenetic trees. *Curr. Biol.* **31**, R59–R64 (2021).

24. P. M. Shih, N. J. Matzke, Primary endosymbiosis events date to the later Proterozoic with cross-calibrated phylogenetic dating of duplicated ATPase proteins. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 12355–12360 (2013).

25. L. W. Parfrey, D. J. G. Lahr, A. H. Knoll, L. A. Katz, Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 13624–13629 (2011).

26. L. Eme, S. C. Sharpe, M. W. Brown, A. J. Roger, On the age of eukaryotes: evaluating evidence from fossils and molecular clocks. *Cold Spring Harb. Perspect. Biol.* **6** (2014).

27. H. C. Betts, M. N. Puttick, J. W. Clark, T. A. Williams, P. C. J. Donoghue, D. Pisani, Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nat Ecol Evol* **2**, 1556–1562 (2018).

28. T. Kleine, H. Palme, K. Mezger, A. N. Halliday, Hf-W chronometry of lunar metals and the age and early differentiation of the Moon. *Science* **310**, 1671–1674 (2005).

29. A. Y. Mulkidjanian, K. S. Makarova, M. Y. Galperin, E. V. Koonin, Inventing the dynamo machine: the evolution of the F-type and V-type ATPases. *Nat. Rev. Microbiol.* **5**, 892–899 (2007).

30. A. G. Stewart, E. M. Laming, M. Sobti, D. Stock, Rotary ATPases--dynamic molecular machines. *Curr. Opin. Struct. Biol.* **25**, 40–48 (2014).

31. N. Iwabe, K. Kuma, M. Hasegawa, S. Osawa, T. Miyata, Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 9355–9359 (1989).

32. J. P. Gogarten, H. Kibak, P. Dittrich, L. Taiz, E. J. Bowman, B. J. Bowman, M. F. Manolson, R. J. Poole, T. Date, T. Oshima, J. Konishi, K. Denda, M. Yoshida, Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 6661–6665 (1989).

33. E. Hilario, J. P. Gogarten, Horizontal transfer of ATPase genes--the tree of life becomes a net of life. *Biosystems.* **31**, 111–119 (1993).

34. R. L. Cross, V. Müller, The evolution of A-, F-, and V-type ATP synthases and ATPases: reversals in function and changes in the H+/ATP coupling ratio. *FEBS Lett.* **576**, 1–4 (2004).

35. A. Y. Mulkidjanian, M. Y. Galperin, K. S. Makarova, Y. I. Wolf, E. V. Koonin, Evolutionary primacy of sodium bioenergetics. *Biol. Direct* **3**, 13 (2008).

36. A. Y. Mulkidjanian, M. Y. Galperin, E. V. Koonin, Co-evolution of primordial membranes and membrane proteins. *Trends Biochem. Sci.* **34**, 206–215 (2009).

37. N. J. Matzke, A. Lin, M. Stone, M. A. B. Baker, Flagellar export apparatus and ATP synthetase: Homology evidenced by synteny predating the Last Universal Common Ancestor. *Bioessays* **43**, e2100004 (2021).

38. V. Müller, G. Grüber, ATP synthases: structure, function and evolution of unique energy converters. *Cell. Mol. Life Sci.* **60**, 474–494 (2003).

39. W. Kühlbrandt, Structure and mechanisms of F-type ATP synthases. *Annu. Rev. Biochem.* **88**, 515–549 (2019).

40. J. P. Gogarten, L. Taiz, Evolution of proton pumping ATPases: Rooting the tree of life. *Photosynth. Res.* **33**, 137–146 (1992).

41. G. Grüber, M. S. S. Manimekalai, F. Mayer, V. Müller, ATP synthases from archaea: the beauty of a molecular motor. *Biochim. Biophys. Acta* **1837**, 940–952 (2014).

42. M. Forgac, Vacuolar ATPases: rotary proton pumps in physiology and pathophysiology. *Nat. Rev. Mol. Cell Biol.* **8**, 917–929 (2007).

43. J. E. Walker, M. Saraste, M. J. Runswick, N. J. Gay, Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* **1**, 945–951 (1982).

44. R. M. Schwartz, M. O. Dayhoff, Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. *Science* **199**, 395–403 (1978).

45. Y. Feng, U. Neri, S. Gosselin, A. S. Louyakis, R. T. Papke, U. Gophna, J. P. Gogarten, The Evolutionary Origins of Extreme Halophilic Archaeal Lineages. *Genome Biol. Evol.* **13** (2021).

46. G. K. A. Hochberg, J. W. Thornton, Reconstructing ancient proteins to understand the causes of structure and function. *Annu. Rev. Biophys.* **46**, 247–269 (2017).

47. M. L. Mascotti, Resurrecting enzymes by ancestral Sequence Reconstruction. *Methods Mol. Biol.* **2397**, 111–136 (2022).

48. P. P. Sharma, W. C. Wheeler, Cross-bracing uncalibrated nodes in molecular dating improves congruence of fossil and molecular age estimates. *Front. Zool.* **11**, 1–13 (2014).

49. G. J. Szöllősi, S. Höhna, T. A. Williams, D. Schrempf, V. Daubin, B. Boussau, Relative time constraints improve molecular dating. (2021). https://doi.org/10.1093/sysbio/syab084.

50. G. J. Szöllosi, B. Boussau, S. S. Abby, E. Tannier, V. Daubin, Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17513–17518 (2012).

51. G. J. Szöllősi, W. Rosikiewicz, B. Boussau, E. Tannier, V. Daubin, Efficient exploration of the space of reconciled gene trees. *Syst. Biol.* **62**, 901–912 (2013).

52. B. Morel, P. Schade, S. Lutteropp, T. A. Williams, G. J. Szöllősi, A. Stamatakis, SpeciesRax: A Tool for Maximum Likelihood Species Tree Inference from Gene Family Trees under Duplication, Transfer, and Loss. *Mol. Biol. Evol.* **39**, msab365 (2022).

53. P. Lapierre, R. Shial, J. P. Gogarten, Distribution of F- and A/V-type ATPases in Thermus scotoductus and other closely related species. *Syst. Appl. Microbiol.* **29**, 15–23 (2006).

54. M. Sumi, M. Yohda, Y. Koga, M. Yoshida, F0F1-ATPase genes from an archaebacterium, Methanosarcina barkeri. *Biochem. Biophys. Res. Commun.* **241**, 427–433 (1997).

55. R. Saum, K. Schlegel, B. Meyer, V. Müller, The F1FO ATP synthase genes in Methanosarcina acetivorans are dispensable for growth and ATP synthesis. *FEMS Microbiol. Lett.* **300**, 230–236 (2009).

56. D. V. Dibrova, M. Y. Galperin, A. Y. Mulkidjanian, Characterization of the N-ATPase, a distinct, laterally transferred Na+-translocating form of the bacterial F-type membrane ATPase. *Bioinformatics* **26**, 1473–1476 (2010).

57. C. J. Castelle, J. F. Banfield, Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. *Cell* **172**, 1181–1197 (2018).

58. C. J. Castelle, C. T. Brown, K. Anantharaman, A. J. Probst, R. H. Huang, J. F. Banfield, Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat. Rev. Microbiol.* **16**, 629–645 (2018).

59. N. Dombrowski, J.-H. Lee, T. A. Williams, P. Offre, A. Spang, Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS Microbiol. Lett.* **366** (2019).

60. B. Wang, W. Qin, Y. Ren, X. Zhou, M.-Y. Jung, P. Han, E. A. Eloe-Fadrosh, M. Li, Y. Zheng, L. Lu, X. Yan, J. Ji, Y. Liu, L. Liu, C. Heiner, R. Hall, W. Martens-Habbena, C. W. Herbold, S.-K. Rhee, D. H. Bartlett, L. Huang, A. E. Ingalls, M. Wagner, D. A. Stahl, Z. Jia, Expansion of Thaumarchaeota habitat range is correlated with horizontal transfer of ATPase operons. *ISME J.* **13**, 3067–3079 (2019).

61. P. N. Evans, D. H. Parks, G. L. Chadwick, S. J. Robbins, V. J. Orphan, S. D. Golding, G. W. Tyson, Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science* **350**, 434–438 (2015).

62. P. N. Evans, J. A. Boyd, A. O. Leu, B. J. Woodcroft, D. H. Parks, P. Hugenholtz, G. W. Tyson, An evolving view of methane metabolism in the Archaea. *Nat. Rev. Microbiol.* **17**, 219–232 (2019).

63. C. W. Stairs, M. M. Leger, A. J. Roger, Diversity and origins of anaerobic metabolism in mitochondria and related organelles. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140326 (2015).

64. R. M. R. Gawryluk, C. W. Stairs, Diversity of electron transport chains in anaerobic protists. *Biochim. Biophys. Acta Bioenerg.* **1862**, 148334 (2021).

65. E. Hilario, J. P. Gogarten, The prokaryote-to-eukaryote transition reflected in the evolution of the V/F/A-ATPase catalytic and proteolipid subunits. *J. Mol. Evol.* **46**, 703–715 (1998).

66. B. Larget, The estimation of tree posterior probabilities using conditional clade probability distributions. *Syst. Biol.* **62**, 501–511 (2013).

67. T. A. Williams, G. J. Szöllősi, A. Spang, P. G. Foster, S. E. Heaps, B. Boussau, T. J. G. Ettema, T. M. Embley, Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E4602–E4611 (2017).

68. T. Cavalier-Smith, Rooting the tree of life by transition analyses. *Biol. Direct* **1**, 19 (2006).

69. J. A. Lake, R. G. Skophammer, C. W. Herbold, J. A. Servin, Genome beginnings: rooting the tree of life. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 2177–2185 (2009).

70. R. Gouy, D. Baurain, H. Philippe, Rooting the tree of life: the phylogenetic jury is still out. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140329 (2015).

71. M. Saraste, P. R. Sibbald, A. Wittinghofer, The P-loop--a common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.* **15**, 430–434 (1990).

72. D. D. Leipe, Y. I. Wolf, E. V. Koonin, L. Aravind, Classification and evolution of P-loop GTPases and related ATPases. *J. Mol. Biol.* **317**, 41–72 (2002).

73. I. B. Schäfer, S. M. Bailer, M. G. Düser, M. Börsch, R. A. Bernal, D. Stock, G. Grüber, Crystal structure of the archaeal A1Ao ATP synthase subunit B from Methanosarcina mazei Gö1: Implications of nucleotide-binding differences in the major A1Ao subunits A and B. *J. Mol. Biol.* **358**, 725–740 (2006).

74. T. A. Mahendrarajah, E. R. R. Moody, D. Schrempf, L. L. Szántho, N. Dombrowski, A. A. Davín, D. Pisani, P. C. J. Donoghue, G. J. Szöllősi, T. A. Williams, A. Spang, ATP synthase evolution on a cross-braced dated tree of life, Zenodo (2023); https://doi.org/10.5281/ZENODO.10012837.

75. R. I. Ponce-Toledo, P. Deschamps, P. López-García, Y. Zivanovic, K. Benzerara, D. Moreira, An Early-Branching Freshwater Cyanobacterium at the Origin of Plastids. *Curr. Biol.* **27**, 386–391 (2017).

76. B. J. Harris, J. W. Clark, D. Schrempf, G. J. Szöllősi, P. C. J. Donoghue, A. M. Hetherington, T. A. Williams, Divergent evolutionary trajectories of bryophytes and tracheophytes from a complex common ancestor of land plants. *Nat Ecol Evol* **6**, 1634–1643 (2022).

77. S. Nelson-Sathi, F. L. Sousa, M. Roettger, N. Lozada-Chávez, T. Thiergart, A. Janssen, D. Bryant, G. Landan, P. Schönheit, B. Siebers, J. O. McInerney, W. F. Martin, Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* **517**, 77–80 (2015).

78. N. Lane, J. F. Allen, W. Martin, How did LUCA make a living? Chemiosmosis in the origin of life. *Bioessays* **32**, 271–280 (2010).

79. V. Sojo, A. Pomiankowski, N. Lane, A bioenergetic basis for membrane divergence in archaea and bacteria. *PLoS Biol.* **12**, e1001926 (2014).

80. Y. Koga, Early evolution of membrane lipids: how did the lipid divide occur? *J. Mol. Evol.* **72**, 274–282 (2011).

81. L. Villanueva, S. Schouten, J. S. S. Damsté, Phylogenomic analysis of lipid biosynthetic genes of Archaea shed light on the 'lipid divide.' *Environ. Microbiol.* **19**, 54–69 (2017).

82. L. Villanueva, F. A. B. von Meijenfeldt, A. B. Westbye, S. Yadav, E. C. Hopmans, B. E. Dutilh, J. S. S. Damsté, Bridging the membrane lipid divide: bacteria of the FCB group superphylum have the potential to synthesize archaeal ether lipids. *ISME J.* **15**, 168–182 (2021).

83. J. Lombard, D. Moreira, Origins and early evolution of the mevalonate pathway of isoprenoid biosynthesis in the three domains of life. *Mol. Biol. Evol.* **28**, 87–99 (2011).

84. Y. Hoshino, E. A. Gaucher, On the Origin of Isoprenoid Biosynthesis. *Mol. Biol. Evol.* **35**, 2185–2197 (2018).

85. J. Lombard, P. López-García, D. Moreira, The early evolution of lipid membranes and the three domains of life. *Nat. Rev. Microbiol.* **10**, 507–515 (2012).

86. S. Mohanty, C. Jobichen, V. P. R. Chichili, A. Velázquez-Campoy, B. C. Low, C. W. V. Hogue, J. Sivaraman, Structural basis for a unique ATP synthase core complex from Nanoarcheaum equitans. *J. Biol. Chem.* **290**, 27280–27296 (2015).

87. S. Krause, A. Bremges, P. C. Münch, A. C. McHardy, J. Gescher, Characterisation of a stable laboratory co-culture of acidophilic nanoorganisms. *Sci. Rep.* **7**, 3289 (2017).

88. B. A. Baker, A. Gutiérrez-Preciado, Á. R. del Río, C. G. P. McCarthy, P. López-García, J. Huerta-Cepas, E. Susko, A. J. Roger, L. Eme, D. Moreira, Several independent adaptations of archaea to hypersaline environments, *bioRxiv* (2023)p. 2023.07.03.547478.

89. P. Narasingarao, S. Podell, J. A. Ugalde, C. Brochier-Armanet, J. B. Emerson, J. J. Brocks, K. B. Heidelberg, J. F. Banfield, E. E. Allen, De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *ISME J.* **6**, 81–93 (2012).

90. K. Andrade, J. Logemann, K. B. Heidelberg, J. B. Emerson, L. R. Comolli, L. A. Hug, A. J. Probst, A. Keillar, B. C. Thomas, C. S. Miller, E. E. Allen, J. W. Moreau, J. J. Brocks, J. F. Banfield, Metagenomic and lipid analyses reveal a diel cycle in a hypersaline microbial ecosystem. *ISME J.* **9**, 2697–2711 (2015).

91. N. Dombrowski, T. A. Williams, J. Sun, B. J. Woodcroft, J.-H. Lee, B. Q. Minh, C. Rinke, A. Spang, Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. *Nat. Commun.* **11**, 3939 (2020).

92. M. Aouad, J.-P. Flandrois, F. Jauffrit, M. Gouy, S. Gribaldo, C. Brochier-Armanet, A divide-and-conquer phylogenomic approach based on character supermatrices resolves early steps in the evolution of the Archaea. *BMC Ecol. Evol.* **22**, 1 (2022).

93. Q. Zhu, U. Mai, W. Pfeiffer, S. Janssen, F. Asnicar, J. G. Sanders, P. Belda-Ferre, G. A. Al-Ghalith, E. Kopylova, D. McDonald, T. Kosciolek, J. B. Yin, S. Huang, N. Salam, J.-Y. Jiao, Z. Wu, Z. Z. Xu, K. Cantrell, Y. Yang, E. Sayyari, M. Rabiee, J. T. Morton, S. Podell, D. Knights, W.-J. Li, C. Huttenhower, N. Segata, L. Smarr, S. Mirarab, R. Knight, Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477 (2019).

94. G. P. Fournier, K. R. Moore, L. T. Rangel, J. G. Payette, L. Momper, T. Bosak, The Archean origin of oxygenic photosynthesis and extant cyanobacterial lineages. *Proc. Biol. Sci.* **288**, 20210675 (2021).

95. K. Mateos, G. Chappell, A. Klos, B. Le, J. Boden, E. Stüeken, R. Anderson, The evolution and spread of sulfur cycling enzymes reflect the redox state of the early Earth. *Sci Adv* **9**, eade4847 (2023).

96. C. Parsons, E. E. Stüeken, C. J. Rosen, K. Mateos, R. E. Anderson, Radiation of nitrogen-metabolizing enzymes across the tree of life tracks environmental transitions in Earth history. *Geobiology* **19**, 18–34 (2021).

97. C. Berney, J. Pawlowski, A molecular timescale for eukaryote evolution recalibrated with the continuous microfossil record. *Proc. Biol. Sci.* **273**, 1867–1872 (2006).

98. D. Chernikova, S. Motamedi, M. Csürös, E. V. Koonin, I. B. Rogozin, A late origin of the extant eukaryotic diversity: divergence time estimates using rare genomic changes. *Biol. Direct* **6**, 26 (2011).

99. J. F. H. Strassert, I. Irisarri, T. A. Williams, F. Burki, A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. *Nat. Commun.* **12**, 1879 (2021).

100. Q. Tang, K. Pang, X. Yuan, S. Xiao, A one-billion-year-old multicellular chlorophyte. *Nat. Ecol. Evol.* **4**, 543–549 (2020).

101. T. M. Gibson, P. M. Shih, V. M. Cumming, W. W. Fischer, P. W. Crockford, M. S. W. Hodgskiss, S. Wörndle, R. A. Creaser, R. H. Rainbird, T. M. Skulski, G. P. Halverson, Precise age of Bangiomorpha pubescens dates the origin of eukaryotic photosynthesis. *Geology* **46**, 135–138 (2018).

102. H. Philippe, P. Forterre, The rooting of the universal tree of life is not reliable. *J. Mol. Evol.* **49**, 509–523 (1999).

103. A. A. Pittis, T. Gabaldón, Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature* **531**, 101–104 (2016).

104. W. F. Martin, M. Roettger, C. Ku, S. G. Garg, S. Nelson-Sathi, G. Landan, Late Mitochondrial Origin Is an Artifact, *Genome biology and evolution*. **9** (2017)pp. 373–379.

105. J. Vosseberg, J. J. E. van Hooff, M. Marcet-Houben, A. van Vlimmeren, L. M. van Wijk, T. Gabaldón, B. Snel, Timing the origin of eukaryotic cellular complexity with ancient duplications. *Nat Ecol Evol* **5**, 92–100 (2021).

106. P. C. J. Donoghue, C. Kay, A. Spang, G. J. Szöllősi, A. Nenarokova, E. R. R. Moody, D. Pisani, T. A. Williams, Defining eukaryotes to dissect eukaryogenesis. *Curr. Biol.* (2023).

107. J. Martijn, J. Vosseberg, L. Guy, P. Offre, T. J. G. Ettema, Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* **557**, 101–105 (2018).

108. S. A. Muñoz-Gómez, E. Susko, K. Williamson, L. Eme, C. H. Slamovits, D. Moreira, P. López-García, A. J. Roger, Site-and-branch-heterogeneous analyses of an expanded dataset favour mitochondria as sister to known Alphaproteobacteria. *Nat Ecol Evol* **6**, 253–262 (2022).

109. K. Katoh, K. Misawa, K.-I. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).

110. A. Criscuolo, S. Gribaldo, BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).

111. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

112. D. R. Utter, X. He, C. M. Cavanaugh, J. S. McLean, B. Bor, The saccharibacterium TM7x elicits differential responses across its host range. *ISME J.* **14**, 3054–3067 (2020).

113. X. He, J. S. McLean, A. Edlund, S. Yooseph, A. P. Hall, S.-Y. Liu, P. C. Dorrestein, E. Esquenazi, R. C. Hunter, G. Cheng, K. E. Nelson, R. Lux, W. Shi, Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 244–249 (2015).

114. T. Aramaki, R. Blanc-Mathieu, H. Endo, K. Ohkubo, M. Kanehisa, S. Goto, H. Ogata, KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).

115. R. L. Tatusov, E. V. Koonin, D. J. Lipman, A genomic perspective on protein families. *Science* **278**, 631–637 (1997).

116. M. Y. Galperin, D. M. Kristensen, K. S. Makarova, Y. I. Wolf, E. V. Koonin, Microbial genome analysis: the COG approach. *Brief. Bioinform.* **20**, 1063–1070 (2019).

117. M. Y. Galperin, Y. I. Wolf, K. S. Makarova, R. Vera Alvarez, D. Landsman, E. V. Koonin, COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.* **49**, D274–D281 (2021).

118. A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Studholme, C. Yeats, S. R. Eddy, The Pfam protein families database. *Nucleic Acids Res.* **32**, D138-41 (2004).

119. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29-37 (2011).

120. S. R. Eddy, A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**, 205–211 (2009).

121. S. R. Eddy, Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).

122. D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

123. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, R. Lanfear, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

124. W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

125. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

126. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).

127. H. Shimodaira, An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508 (2002).

128. UniProt Consortium, UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

129. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

130. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

131. A. L. Mitchell, T. K. Attwood, P. C. Babbitt, M. Blum, P. Bork, A. Bridge, S. D. Brown, H.-Y. Chang, S. El-Gebali, M. I. Fraser, J. Gough, D. R. Haft, H. Huang, I. Letunic, R. Lopez, A. Luciani, F. Madeira, A. Marchler-Bauer, H. Mi, D. A. Natale, M. Necci, G. Nuka, C. Orengo, A. P. Pandurangan, T. Paysan-Lafosse, S. Pesseat, S. C. Potter, M. A. Qureshi, N. D. Rawlings, N. Redaschi, L. J. Richardson, C. Rivoire, G. A. Salazar, A. Sangrador-Vegas, C. J. A. Sigrist, I. Sillitoe, G. G. Sutton, N. Thanki, P. D. Thomas, S. C. E. Tosatto, S.-Y. Yong, R. D. Finn, InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **47**, D351–D360 (2019).

132. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

133. J. Bergsten, A review of long-branch attraction. *Cladistics* **21**, 163–193 (2005).

134. A. M. Waterhouse, J. B. Procter, D. M. A. Martin, M. Clamp, G. J. Barton, Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).

135. G. E. Crooks, G. Hon, J.-M. Chandonia, S. E. Brenner, WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).

136. F. Burki, A. J. Roger, M. W. Brown, A. G. B. Simpson, The new tree of eukaryotes. *Trends Ecol. Evol.* **35**, 43–55 (2020).

137. L. Eme, D. Tamarit, E. F. Caceres, C. W. Stairs, V. De Anda, M. E. Schön, K. W. Seitz, N. Dombrowski, W. H. Lewis, F. Homa, J. H. Saw, J. Lombard, T. Nunoura, W.-J. Li, Z.-S. Hua, L.-X. Chen, J. F. Banfield, E. S. John, A.-L. Reysenbach, M. B. Stott, A. Schramm, K. U. Kjeldsen, A. P. Teske, B. J. Baker, T. J. G. Ettema, Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature* **618**, 992–999 (2023).

138. J. F. Parham, P. C. J. Donoghue, C. J. Bell, T. D. Calway, J. J. Head, P. A. Holroyd, J. G. Inoue, R. B. Irmis, W. G. Joyce, D. T. Ksepka, J. S. L. Patané, N. D. Smith, J. E. Tarver, M. van Tuinen, Z. Yang, K. D. Angielczyk, J. M. Greenwood, C. A. Hipsley, L. Jacobs, P. J. Makovicky, J. Müller, K. T. Smith, J. M. Theodor, R. C. M. Warnock, M. J. Benton, Best practices for justifying fossil calibrations. *Syst. Biol.* **61**, 346–359 (2012).

139. H. Shimodaira, M. Hasegawa, CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**, 1246–1247 (2001).

CHAPTER 4

A rooted phylogeny resolves early bacterial evolution

Gareth A. Coleman*, Adrián A. Davín*, Tara A. Mahendrarajah, Lénárd L. Szánthó, Anja Spang, Philip Hugenholtz**, Gergely J. Szöllősi**, Tom A. Williams**
*these authors contributed equally to this work
**these authors contributed equally to this work

Science, 2021

SUMMARY AND CONTRIBUTIONS

This project was a massive collaborative effort requiring the expertise of members from four research groups across three time zones during the COVID-19 pandemic lockdown period. The goal of this project was to examine bacterial evolution, namely to infer a rooted bacterial phylogeny without including an archaeal outgroup, to assess the verticality of bacterial evolution, and examine the nature of the last bacterial common ancestor (LBCA). Gene tree-species tree reconciliation implemented with Amalgamated Likelihood Estimation (ALE) was used to model the evolutionary histories of 11,272 gene families from all bacteria in this study, against a species tree of those same bacteria. ALE quantifies instances of gene duplications, transfers, originations, and losses (DTL), and assigns a probability (posterior probability, PP) for genes families across the species phylogeny. Different root positions were statistically tested considering the gene families' evolutionary histories along the species tree, resulting in a statistically favored root position between the two major bacterial clades, Terrabacteria and Gracilicutes, and that the candidate phyla radiation (CPR) share a common ancestor with the Chloroflexota. Both findings challenge proposals that the root of the bacterial tree sits between the CPR and all other Bacteria, and instead highlights that the CPR likely evolved via genome reduction from a free-living ancestor and that their compositional biases result in phylogenetic artifacts in traditional outgroup rooting methods. Modeling the duplications, transfers, originations, and losses also provides the foundation for quantifying the degree of vertical and horizontal gene flow acting upon bacterial evolutions. Results indicated that the majority of genes (92%) experience some level of horizontal transfer during their evolutionary history, however 66% of all genes were transmitted vertically (and 34% horizontally), which together make a tree a suitable representation of bacterial evolution. The PP values assigned to COG families were used to determine the presence or absence of genes and metabolic pathways in LBCA. We inferred LBCA to be an already complex rod-shaped diderm, with flagella, pili, and lipopolysaccharides. While several carbohydrate metabolizing pathways had high support, carbon fixation pathways were more patchily recovered. We resolved components of the Wood-Ljungdahl pathway, the tricarboxylic acid (TCA) cycle, and the pentose phosphate pathway (PPP), but did not recover key enzymes and the directionality of some reactions is debated. However, there was high support for components of the Rhodobacter nitrogen-fixing (Rnf) complex, which together with the inferred methyl branch of the WLP pathway could hint at possible acetogenic growth in LBCA. Finally, we found a near-complete CRISPR-Cas system, implying that LBCA was likely exposed to viruses and other parasitic replicators.

Phylogenomic analyses, including the outgroup-free rooting and measure of verticality, were primarily conducted by Gareth A. Coleman, Adrián A. Davín, Lénárd L. Szánthó, Philip Hugenholtz, Gergely J. Szöllősi, and Tom A. Williams, with Gergely J. Szöllősi being the primary contributor of new analytical techniques. The metabolic reconstruction was performed by myself, Gareth A. Coleman, Adrián A. Davín, and Anja Spang. Data analysis and writing sessions were several hours long and occurred 1-2 times per week with all authors. I participated in

additional weekly meetings with Gareth A. Coleman and Anja Spang to analyze the gene content and metabolic profile of the last bacterial common ancestor (LBCA).

My specific methodological contributions include the following: I annotated all gene sequences comprising the clustered gene families used in the probabilistic gene tree-species tree reconciliations. I assessed the consistency of different annotation profiles (i.e., KOs vs COGs) in coverage of the assigned COG family and filtered out families that did not have at least 50% coverage by a single KO congruent with the COG family identity. I compiled all relevant COG families, their descriptions, categories, corresponding KOs, KO descriptions, and assigned posterior probabilities (PP) across all nodes and tips in the concatenated species tree, from two analyses (focal and secondary), into large data tables for parsing. My metabolic reconstruction analysis focused on nine ancestral node positions: Root 1, Root 2, Root 3, Terrabacteria + DST, Terrabacteria, Gracilicutes, CPR + Chloroflexota, Chloroflexota, and CPR. Together with Gareth A. Coleman and Anja Spang, we formulated a procedure from which to assess the presence or absence of key gene families in the reconstruction. We assigned three PP ranges including high (PP>0.95), moderate (PP=0.75-0.95), and low (PP=0.50-0.75), with anything below 0.50 being considered absent. Across the focal and secondary datasets, I identified COGs (and their corresponding KOs) affiliated with key metabolic and biosynthetic pathways, including: carbohydrate metabolism, carbon metabolism, cell structure proteins, motility and chemotaxis, lipid biosynthesis, cell division, energy generation, and defense mechanisms, among others.

I parsed the table of all genes assessed in the reconciliations and assessed their presence/absence based on the PP values or combination of overall PP values for gene families associated with enzymatic complexes or pathways. I carefully inspected the reconciled genes to develop a profile of LBCA's carbon metabolism, autotrophy, respiratory complexes, and defense mechanisms. I assisted with the construction of the KEGG metabolic pathways presence/absence diagram in fig. S17 (see supplementary information). In addition, I designed a heatmap to visualize the percentage distribution of COGs per clade in the rooted concatenated species tree and across the nine highlighted node positions (see fig. S18). I designed and built the main text Fig. 4 (ancestral reconstruction of LBCA) with Gareth A. Coleman and Anja Spang. I finalized all PP values across all visual reconstructions (Fig. 4, fig. S17) and all texts (main, supplemental, and additional). I contributed to data interpretation of the metabolic inference but also the phylogenomic results, visualization, and writing and revision of the original and final versions of the manuscript.

RESEARCH ARTICLE SUMMARY

STRUCTURED ABSTRACT

INTRODUCTION

Bacteria are the most diverse and abundant cellular organisms on Earth, and in recent years environmental genomics has revealed the existence of an enormous diversity of previously unknown lineages. Despite the abundance of genomic sequence data, the root of the bacterial tree and the nature of the most recent common ancestor of Bacteria have remained elusive. The problem is that even with the help of new data, tracing billions of years of bacterial evolution back to the root has remained challenging because standard phylogenetic models do not account for the full range of evolutionary processes that shape bacterial genomes. In particular, standard models treat horizontal gene transfer as an impediment to the reconstruction of the tree of life that must be removed from analyses. But if horizontal gene transfer is modeled appropriately, it can provide information about the deep past that is unavailable to standard methods.

RATIONALE

We reconstructed and rooted the bacterial tree by applying a hierarchical phylogenomic approach that explicitly uses information from gene duplications and losses within a genome as well as gene transfers between genomes. This approach allowed us to root the tree without including an archaeal outgroup. Outgroup-free rooting is a promising approach for Bacteria, both because the position of the universal root is uncertain and because the long branch separating Bacteria from Archaea has the potential to distort the reconstruction of within-Bacteria relationships. Outgroup-free gene tree-species tree reconciliation allowed us to quantitatively model both the vertical and horizontal components of bacterial evolution and integrate information from 11,272 gene families to resolve the root of the bacterial tree. Notably, these analyses also provided estimates of the gene content of the last bacterial common ancestor.

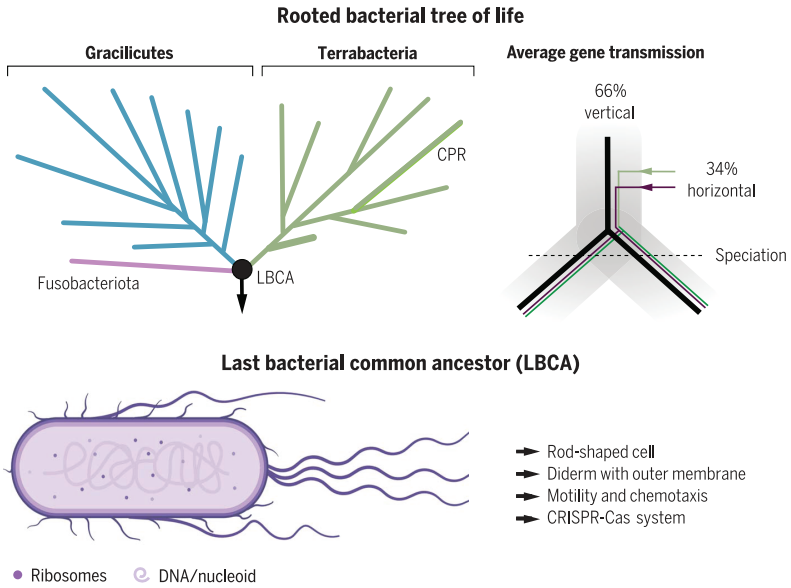
RESULTS

Our analyses place the root between two major bacterial clades, the Gracilicutes and Terrabacteria. We found no support for a root between the Candidate Phyla Radiation (CPR), a lineage comprising putative symbionts and parasites with small genomes, and all other Bacteria. Instead, the CPR was inferred to be a member of the Terrabacteria and formed a sister lineage to the Chloroflexota and Dormibacterota. This suggests that the CPR evolved by reductive genome evolution from free-living ancestors. Gene families inferred to have been present at the root indicate that the last bacterial common ancestor was already a complex double-membraned cell capable of motility and chemotaxis that possessed a CRISPR-Cas system. Although ~92% of gene families have experienced horizontal transfers during their history, tracing their evolution along the most likely rooted tree revealed that about two-thirds of gene transmissions have been vertical. Thus, bacterial evolution has a

major vertical component, despite a profound impact of horizontal gene transfer through time. Horizontal gene flows can also provide insight into the temporal sequence of events during bacterial diversification, because donor lineages must be at least as old as recipients. Analysis of gene transfers in our dataset suggests that the diversification of the Firmicutes, CPR, Acidobacteriota, and Proteobacteria is the oldest among extant bacterial phyla.

CONCLUSION

The vertical and horizontal components of genome evolution provide complementary sources of information about bacterial phylogeny. The vertical component provides a robust framework for interpreting species diversity and allows us to reconstruct ancestral states, while the horizontal component helps to root the vertical tree and orient it in time. The inferred Gracilicutes-Terrabacteria root will be useful for investigating the tempo and mode of bacterial diversification, metabolic innovation, and changes in cell architecture such as the evolutionary transitions between double (diderm) and single (monoderm) membranes. Future development of methods that harness the complementarity of vertical and horizontal gene transmission will continue to further our understanding of life on Earth.



A rooted phylogeny of Bacteria. The reconciliation of bacterial gene phylogenies places the root between the major clades of Gracilicutes (including Proteobacteria and Bacteroidota) and Terrabacteria (including Firmicutes and Cyanobacteria). On the basis of this hypothesis, ancestral genome reconstruction predicts that the last bacterial common ancestor (LBCA) was a complex, double-membraned cell and that, on average, two-thirds of gene transmissions have been vertically inherited along this rooted tree.

ABSTRACT

A rooted bacterial tree is necessary to understand early evolution, but the position of the root is contested. Here, we model the evolution of 11,272 gene families to identify the root, extent of horizontal gene transfer (HGT), and the nature of the last bacterial common ancestor (LBCA). Our analyses root the tree between the major clades Terrabacteria and Gracilicutes and suggest that LBCA was a free-living flagellated, rod-shaped double-membraned organism. Contrary to recent proposals, our analyses reject a basal placement of the Candidate Phyla Radiation, which instead branches sister to Chloroflexota within Terrabacteria. While most gene families (92%) have evidence of HGT, overall, two-thirds of gene transmissions have been vertical, suggesting that a rooted tree provides a meaningful frame of reference for interpreting bacterial evolution.

A species tree captures the relationships among organisms but requires a root to provide the direction of evolution. Rooting deep radiations (*1*) is among the greatest challenges in phylogenetics, and there is no consensus on the root of the bacterial tree. On the basis of evidence (*2*–*5*)) that the root of the tree of life lies between Bacteria and Archaea, early analyses with an archaeal outgroup placed the bacterial root near Aquificales and Thermotogales (*6*, *7*) or Planctomycetes (*8*). Alternative approaches, including analyses of gene flows and the evolution of multimeric protein complexes as well as other complex characters (*9*), have instead suggested roots within the monoderm (single-membrane) Bacteria (*10*) or between Chloroflexi and all other cellular life (*9*). The development of techniques for sequencing microbes directly from environmental samples, without the need for laboratory cultivation, has greatly expanded the genomic representation of natural prokaryotic diversity (*11*–*14*). Recent phylogenomic analyses of expanded sets of diverse bacteria have placed the root between one of the recently identified groups, the Candidate Phyla Radiation [CPR, also known as Patescibacteria (*15*, *16*)] and all other Bacteria (*11*, *16*, *17*). The CPR is characterized by small cells and genomes that appear to have predominantly symbiotic or parasitic lifestyles, but much remains to be learned about their ecology and physiology (*15*, *17*–*19*)). If correct, the early divergence of the CPR has important implications for our understanding of the earliest period of cellular evolution. Along with evidence that the root of the archaeal domain lies between the reduced and predominantly host-associated DPANN superphylum (originally named after Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, and Nanohaloarchaeota) and all other Archaea (*1*, *20*), the CPR root implies that streamlined, metabolically minimalist prokaryotes have coexisted with the more familiar, self-sufficient lineages throughout the history of cellular life (*19*, *21*).

Improved taxon sampling can help to resolve evolutionary relationships (*22*, *23*), and the quantity and diversity of genome sequence data now available presents an opportunity to address long-standing questions about the origins and diversification of Bacteria. However, deep phylogenetic divergences are difficult to resolve, both because the phylogenetic signal for such relationships is overwritten by new changes over time, and also because the process of sequence evolution is more complex than the best-fitting models currently available. In particular, variation in nucleotide or amino acid composition across the sites of the alignment and the branches of the tree can induce long branch attraction (LBA) artifacts in which deep-branching, fast-evolving, poorly sampled or compositionally biased lineages group together irrespective of their evolutionary history (*24*). These issues are widely appreciated (*11*) but are challenging to address adequately, particularly when sequences from thousands of taxa (*11*, *13*, *14*, *16*, *17*) are used to estimate trees of global prokaryotic diversity, which precludes the use of the best-fitting phylogenetic methods available.

ARCHAEAL OUTGROUP ROOTING DOES NOT UNAMBIGUOUSLY ESTABLISH THE ROOT OF THE BACTERIAL TREE

The standard approach to rooting is to include an outgroup in the analysis, and all published phylogenies in which CPR forms a sister lineage to the rest of the Bacteria (*11*, *16*, *17*) have made use of an archaeal outgroup. Outgroup rooting on the bacterial tree, however, has three serious limitations. First, interpretation of the results requires the assumption that the root of the tree of life lies between Bacteria and Archaea. While this is certainly the consensus view, the available evidence is limited and difficult to interpret (*2*–*5*, *25*), and alternative hypotheses in which the universal root is placed within Bacteria have been proposed on the basis of indels (*26*, *27*) or the analysis of slow-evolving characters (*9*). Second, the long branch leading to the archaeal outgroup has the potential to distort within-Bacteria relationships because of LBA. Third, joint analyses of Archaea and Bacteria use a smaller number of genes that are widely conserved and have evolved vertically since the divergence of the two lineages, and sequence alignment is more difficult owing to the low sequence identity between homologs of the two domains.

We evaluated the performance of outgroup rooting on a bacterial tree using 265 Bacteria (see below) and 149 Archaea from a shared subset of 29 phylogenetic markers (table S1). Using this archaeal outgroup, the maximum likelihood (ML) phylogeny under the best-fitting model (LG+C60+R8+F, which accounts for site heterogeneity in the substitution process) placed the bacterial root between a clade comprising Cyanobacteria, Margulisbacteria, CPR, Chloroflexota, and Dormibacterota on one side of the root and all other taxa on the other (fig. S1). However, bootstrap support for this root, and indeed many other deep branches in both the bacterial and archaeal subtrees, was low (50 to 80%). We therefore used approximately unbiased (AU) tests (*28*) to determine whether a range of published alternative rooting hypotheses (table S2) could be rejected, given the model and data. The AU test asks whether the optimal trees that are consistent with these other hypotheses have a significantly worse likelihood score than the ML tree. In this case, the likelihoods of all tested trees were statistically indistinguishable (AU test, $P > 0.05$, table S2), indicating that outgroup rooting cannot resolve the bacterial root on this alignment.

AN ALTERNATIVE TO OUTGROUP ROOTING FOR DEEP MICROBIAL PHYLOGENY

Given the limitations of using a remote archaeal outgroup to establish the root of the bacterial tree, we explored outgroup-free rooting using gene tree-species tree reconciliation (*1*, *29*–*31*). We recently applied this approach to root the archaeal tree (*1*), and similar approaches have been used to investigate the root of eukaryotes (*32*, *33*) and to map and characterize whole-genome duplications in plants (*34*). Gene tree-species tree reconciliation methods work by adding a layer to the standard framework for inferring trees from molecular data. This additional step models the way in which gene trees can differ from each other and the overarching rooted species tree. Substitution models [such as LG (*35*)] describe how the constituent sequences of a gene family evolve along a gene tree via a series of amino acid substitutions that allow us to

infer the most likely gene tree. Reconciliation models describe how a gene tree evolves along the rooted species tree, beginning with gene birth (origination) and followed by a combination of vertical descent and events such as gene duplications, transfers, and losses (this series of events is known as a DTL reconciliation). Combining the substitution-based modeling of sequences along the gene tree with the reconciliation-based modeling of gene trees along a rooted species tree allows us to infer the most likely rooted species tree from the constituent gene families. In other words, reconciliation methods aggregate phylogenetic signal across gene families and, because the likelihood of reconciliations depends on the position of the root, can be used to test the support for competing root positions (*1*, *29*), providing a genome-wide (and gene transfer–aware) extension of the classical approach used to root the tree of life on the basis of ancient gene duplications (*3*, *4*).

Our method, amalgamated likelihood estimation (ALE), improves on earlier approaches by explicitly accounting for uncertainty in the gene tree topologies and in the events leading to those topologies while simultaneously estimating rates of gene duplication, transfer, and loss directly from the data (*31*). Simulations suggest that root inferences under ALE are robust to variation in taxon sampling and the proportion of extinct lineages (fig. S2), that the method finds the correct root even under high levels of gene transfer (*1*, *29*), and that the numbers of D, T, and L events are accurately recovered from the data (figs. S3 to S8). These results suggest that ALE is appropriate for the problem at hand (*36*).

ROOTING BACTERIA WITHOUT AN OUTGROUP

To obtain an unrooted species tree for ALE analysis, we selected a focal dataset of 265 genomes representative of bacterial diversity according to the Genome Taxonomy Database (GTDB) (*13*). We inferred the tree from a concatenation of 62 conserved single-copy markers (table S1) using the LG+C60+R8+F model in IQ-Tree 1.6.10 (Fig. 1), which was chosen as the best-fitting model using the Bayesian information criterion (BIC) (*37*). This yielded highly congruent trees when removing 20 to 80% of the most compositionally heterogeneous sites from the alignment (fig. S9), suggesting that the key features of the topology are not composition-driven LBA artifacts. One exception was the position of the Fusobacteriota, which was recovered as a sister lineage to a clade comprising Deinococcota, Synergistota, and Thermotogota (DST) when 20% of the most heterogeneous sites were removed (fig. S9A) but was recovered as a single lineage between Terrabacteria plus DST and Gracilicutes in all other trees.

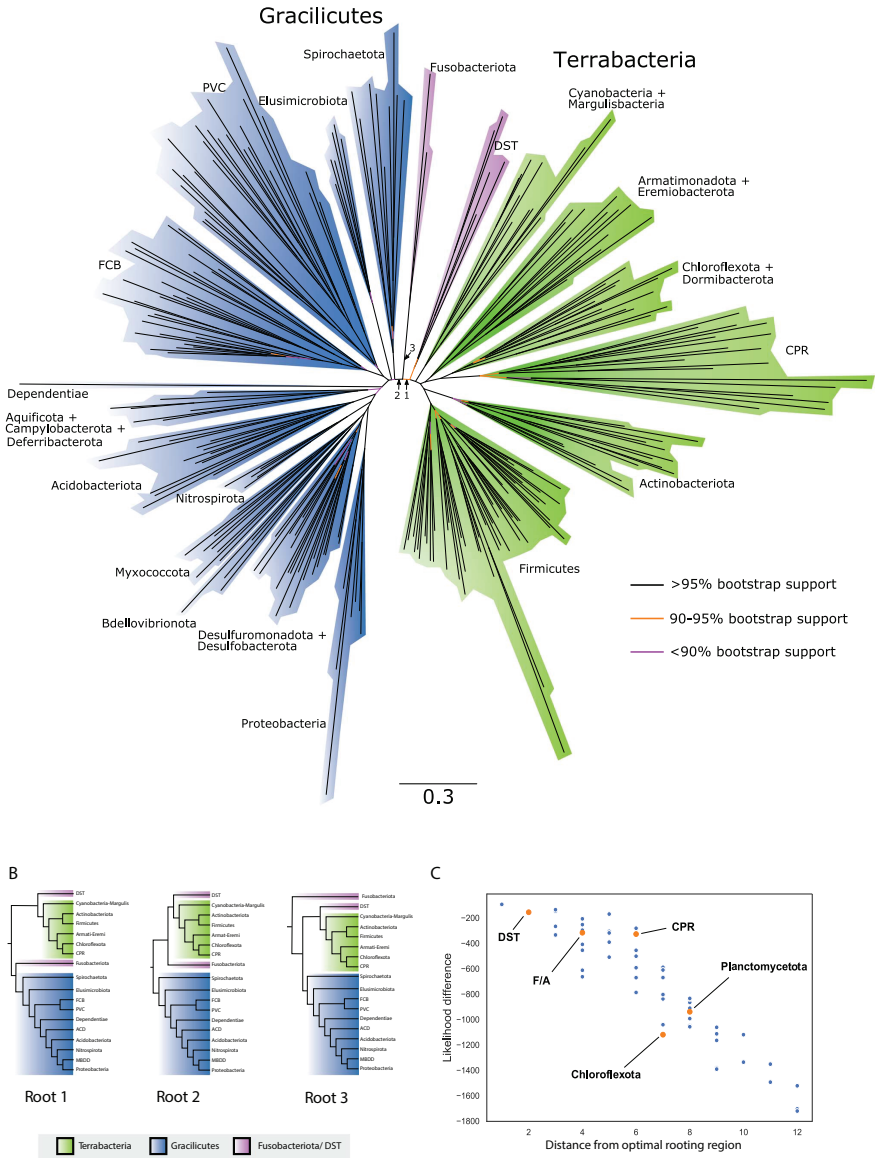


Fig. 1 | A rooted phylogeny of Bacteria. (**A**) We used gene tree-species tree reconciliation to infer the root of the bacterial tree. The unrooted maximum likelihood phylogeny was inferred from a concatenation of 62 marker genes under the best-fitting model, LG+C60+R8+F, which accounts for site heterogeneity in the substitution process and uses a mixture of eight substitution rates estimated from the data to model across-site evolutionary rate variation. Branches are colored according to bootstrap support value. The root falls between two major clades of Bacteria, the Gracilicutes and the Terrabacteria, on one of three statistically equivalent adjacent branches indicated by arrows, shown as rooted trees in (**B**). All alternative roots tested were rejected (tables S3 and S4), with likelihoods decreasing as a function of distance from the root region, as shown in (**C**). Previously proposed root positions, including the CPR root, are highlighted in red. FCB are the Fibrobacterota, Chlorobia, Bacteroidota, and related lineages; PVC are the Planctomycetota, Verrucomicrobiota, Chlamydiota, and related lineages; DST are the Deinococcota,

Synergistota, and Thermotogota; ACD are Aquificota, Campylobacterota, and Deferribacterota; F/A are Firmicutes and Actinobacteriota; MBDD are Myxococcota, Bdellovibrionota, Desulfomonadota, and Desulfobacterota. Scale bar, 0.3 amino acid substitutions per site.

We used ALE to test the support for 62 root positions (tables S3 and S4) on the unrooted topology by reconciling gene trees for 11,272 homologous gene families [inferred using MCL (*38*)] from the 265 bacterial genomes. Note that this method does not assume that the root lies between Bacteria and Archaea. In addition to testing root positions corresponding to published hypotheses, we exhaustively tested all inner nodes of the tree above the phylum level. The ALE analysis rejected all of the root positions tested (AU test, $P < 0.05$) except for three adjacent branches, lying between the two major clades of Gracilicutes (comprising most diderm lineages) and Terrabacteria (comprising monoderm and atypical diderm lineages) (Fig. 1); the difference between the three root positions was the position of the Fusobacteriota in relation to these two major clades (Fig. 1B). Alternative roots were rejected with increasing confidence as distance from the optimal root region increased (Fig. 1C and table S3).

We tested the robustness of the inferred root region by (i) excluding gene families with extreme duplication, transfer, or loss rates; (ii) repeating the analysis using gene families constructed with an assignment to families in the Clusters of Orthologous Genes (COG) (*39*) ontology; and (iii) repeating the analysis on a secondary independent sampling of the tree, in which CPR makes up 40% of the genomes (*11*) (figs. S10 to S13 and table S5). These analyses consistently recovered the root between the Gracilicutes and Terrabacteria, regardless of the position of the Fusobacteriota. A Gracilicutes-Terrabacteria root was previously reported (*40*, *41*), but these studies did not include the CPR, which has recently been suggested to represent the earliest diverging bacterial lineage (*11*, *16*). Our outgroup-free analysis consistently recovered CPR nested within the Terrabacteria, as a sister clade to Chloroflexota and Dormibacterota, even with CPR representing more than 40% of the taxa included. This finding implies that the CPR evolved by genome reduction from a free-living ancestor, a scenario that has been proposed previously (*21*).

Transfers contain information about the relative timing of divergences, because for each transfer, the donor must be at least as old as the recipient (*42*, *43*). To establish the relative ages of the crown groups of different phyla, we used high-confidence relative age constraints recovered in at least 95 of 100 bootstrap replicates common to the focal and secondary datasets (*36*). Simulations suggest that this approach accurately recovers relative clade ages (98.4% accuracy on a simulated dataset the same size as the focal dataset, fig. S14). Our analysis (Fig. 2) predicts that the Firmicutes crown group is the oldest among extant bacterial phyla (median rank: 2 ± 1.43 SD) followed by the crown groups of the CPR (median rank: 3 ± 2), Proteobacteria (median rank: 3 ± 1.59), and Acidobacteriota (median rank: 3 ± 1.56), suggesting that these lineages were the earliest to diversify within Bacteria. The crown groups of lineages predominantly associated with animal hosts, Spirochaetota (median rank: 10 ± 0.85) and Elusimicrobiota (median rank: 11 ± 0.62), appear to be the youngest among extant phyla.

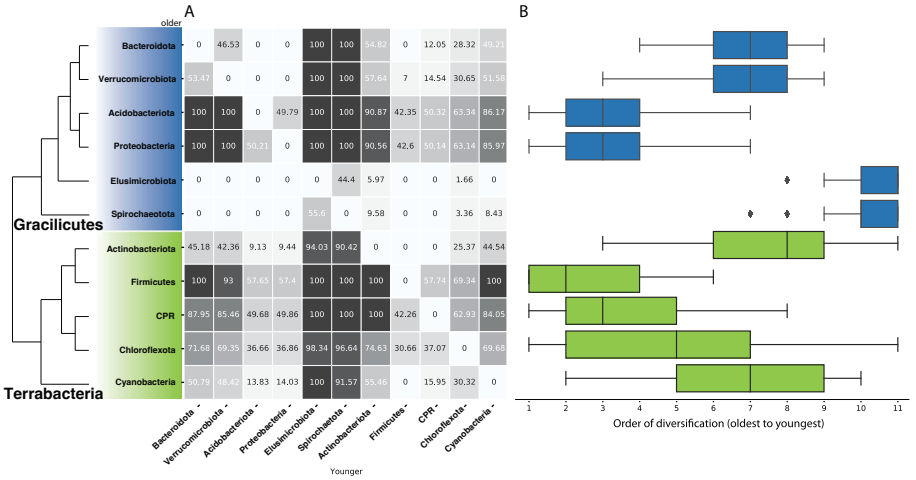


Fig. 2 | Relative crown group ages of major bacterial phyla. Gene transfers that occurred during the diversification of Bacteria provide a record of the temporal sequence of events. We used the information provided by directional (donor-to-recipient) patterns of gene transfer to infer the relative ages of bacterial crown groups, focusing on phyla represented by at least five genomes in both of our datasets. To summarize this time information, we sampled 1000 time orders that were fully compatible with the constraints recovered from both datasets. (**A**) Pairwise relative ages of phyla. The proportion of sampled time orders in which each phylum on the *x* axis was recovered as younger than each phylum on the *y* axis. (**B**) Relative age distributions of major phyla. For each sampled time order, we ranked the phyla from oldest (1) to youngest (11) and plotted the distribution of the ranks. The crown group radiations of Firmicutes, CPR, Proteobacteria, and Acidobacteriota appear to be the oldest among sampled phyla, while those of Elusimicrobiota and Spirochaetota are the youngest.

IS BACTERIAL EVOLUTION TREELIKE?

How much of bacterial evolution can be explained by the concept of a rooted species tree? Horizontal gene transfer (HGT) is frequent in prokaryotes, and published analyses indicate that most or all prokaryotic gene families have experienced HGT during their history (*1*, *44*). This implies that there is no single tree that fully describes the evolution of all bacterial genes or genomes (*45*, *46*). Extensive HGT is existentially challenging for concatenation, because it greatly curtails the number of genes that evolve on a single underlying tree (*47*). Phylogenetic networks (*46*, *48*) were the first methods to explicitly acknowledge nonvertical evolution, but they can be difficult to interpret biologically. Gene tree-species tree reconciliation unites tree and network-based approaches by modeling both the horizontal components of genome evolution (a fully reticulated network allowing all possible transfers) and the vertical trace (a common rooted species tree). This framework enables us to quantify the contributions of vertical and horizontal processes to bacterial evolutionary history.

Our analyses (Fig. 3) reveal that most bacterial gene families present in two or more species (9678 of 10,518 MCL families, or 92%) have experienced at least one gene transfer during their evolution; only very small families have escaped transfer entirely on the time scales considered here (fig. S15). Consistent with previous analyses (*1*, *49*), transfer rates vary across gene

functional categories, with genes encoding proteins involved in defense mechanisms (such as antibiotic biosynthesis) and in the production of secondary metabolites being the most frequently transferred, and those coding for translational and cell cycle proteins the least (Fig. 3B). Despite this accumulation of HGT, most gene families evolve vertically the majority of the time, with mean verticality estimated to be 64% in the focal and 68% in the secondary dataset.

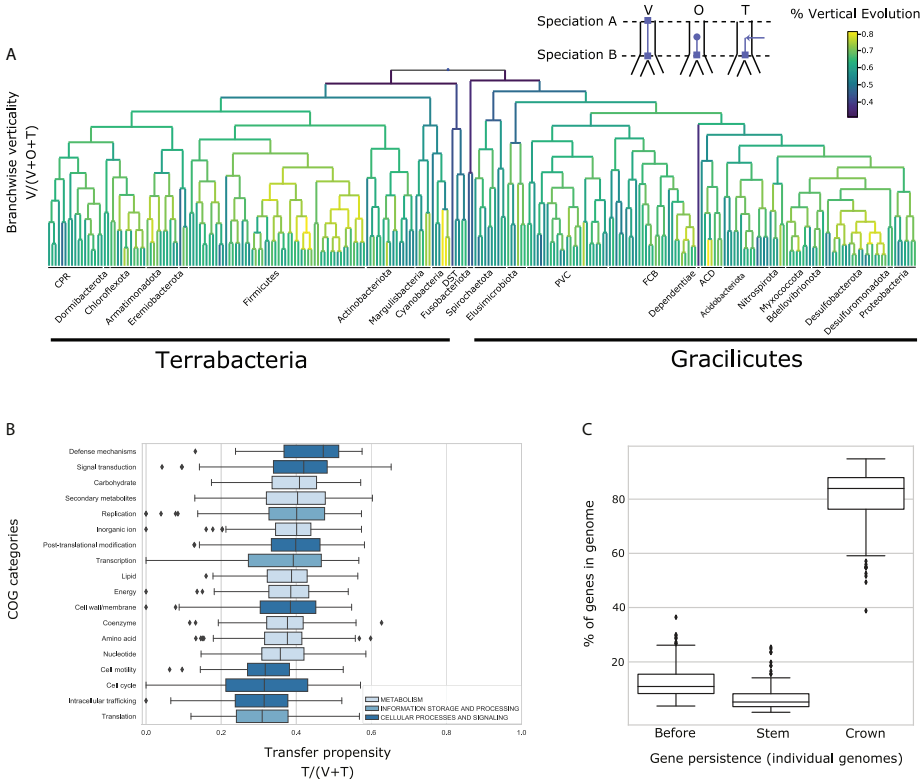


Fig. 3 | The verticality of bacterial genome evolution. (**A**) The rooted bacterial species tree (Fig. 1), with branches colored according to verticality: the fraction of genes at the bottom of a branch that descend vertically from the top of that branch (see inset; V, vertical; O, origination; T, transfer into a branch) (*36*). Node heights reflect relative time order consistent with highly supported gene transfers (Fig. 2). (**B**) Transfer propensity by COG functional category; that is, the proportion of gene tree branches that are horizontal T/(V+T) for COG gene families. Genes involved in information processing, particularly translation (J), show the lowest transfer propensity (median: 0.31), while genes involved in cell defense mechanisms (V, such as genes involved in antibiotic defense and biosynthesis) are most frequently transferred (median transfer propensity: 0.47). (**C**) From the genome's eye view, this combination of vertical and horizontal processes gives rise to a distribution of gene persistences (residence times), reflecting the point in evolutionary history [within the Crown group, on the Stem, or earlier (Before)] at which the gene was most recently acquired. Across all phyla examined, 82% of genes on sampled genomes were most recently acquired since the crown group radiation of that phylum. Lineage acronyms are as in Fig. 1.

Genome-wide reconciliation of gene trees with the species tree demonstrates that the optimal rooted species tree provides an apt summary of much of bacterial evolutionary history, even for the deepest branches of the tree (*50*). From the gene's eye view, gene families evolve neither entirely vertically nor horizontally; core genes are occasionally transferred, and even frequently exchanged genes contribute useful vertical signal; for example, the median number of genes that evolve vertically on a branch of the species tree is 998.92 in the focal analysis (table S6), far greater than the number of genes that have been concatenated at the level of all Bacteria. From the perspective of the genome, constituent genes have different ages (or residence times), corresponding to the time at which they originated or were most recently acquired by gene transfer, within the resolution of our taxonomic sampling.

This analysis indicates that, on average, 82% of all genes from adequately represented phyla (five or more genomes) were most recently acquired after the diversification of that phylum, although all genomes retain a smaller proportion (10 to 27%) of genes that have descended vertically from the stem lineage of their phylum or even earlier (Fig. 3C). There are two explanations for this distribution of gene persistence times: (i) de novo gene origination within phyla (that is, lineage-specific gene families) and (ii) the cumulative impact of gene transfer, which curtails gene persistence times when looking back from the present day even though most transmissions are vertical.

ANCESTRAL PROTEOME OF THE LAST BACTERIAL COMMON ANCESTOR

Reconciliation analyses not only allow us to infer the acquisition of genes across the tree but also to estimate the metabolic potential of the last bacterial common ancestor (LBCA). We built a second, smaller set of COG-based gene families better suited for functional annotation and reconciled their gene trees with the species tree (*36*). In the following reconstruction, we indicate when gene content inferences differ between roots (*36*). Posterior probabilities (PPs) for genes directly relevant to our reconstruction are provided in table S7, and all of the pathways we discuss below were confirmed in our analysis of the secondary dataset (*36*). From the root placement and estimated rates of gene family extinction in the focal analysis (*1*), we predict that LBCA encoded 1293 to 2143 COG family members, the majority of which (median estimates: 65 to 69.5%; 95% confidence interval: 57 to 82%) survived to be sampled in at least one present-day genome. On the basis of the relationship between COG family members and genome size for extant Bacteria (Pearson's correlation coefficient = 0.96, $P = 8 \times 10^{-153}$), we estimate the genome size of LBCA to be 2.7 ± 0.4 Mb (SE) for root 1 of the focal analysis (Fusobacteriota with Terrabacteria) (Fig. 1B), 2.6 ± 0.4 Mb for root 2 (Fusobacteriota with Gracilicutes), and 1.6 ± 0.5 Mb for root 3 (Fusobacteriota root). Under all three roots, the trend in genome size evolution from LBCA to modern taxa is an ongoing moderate increase through time in estimated COG family complements and genome sizes. The most notable departure from this trend is a reduction in genome size of between 0.47 and 0.56 Mb on the CPR stem

lineage after divergence from their common ancestor with Chloroflexota and Dormibacterota (fig. S16). COG families lost on the CPR stem include components of the electron transport chain, carbon metabolism, flagellar biosynthesis and motor switch proteins, amino acid biosynthesis, the Clp protease subunit ClpX, and RNA polymerase sigma factor-54 (table S8), consistent with their absence in extant CPR (*18*).

The inferred ancestral gene set for LBCA includes most components of the modern bacterial transcription, translation, and DNA replication systems (table S7). This gene set also includes an FtsZ-based cell division machinery and pathways for signal transduction, membrane transport, and secretion (Fig. 4) (*36*). Further, we identified proteins involved in bacterial phospholipid biosynthesis, suggesting that LBCA had bacterial-type ester-lipid membranes (Fig. 4). We also identified most of the proteins required for flagella and pili synthesis and those for quorum sensing, suggesting that LBCA was motile (*51*, *52*). Given that bacterial genes are typically maintained by strong purifying selection (*53*), these findings imply that LBCA lived in an environment in which dispersal, chemotaxis, and surface attachment were advantageous.

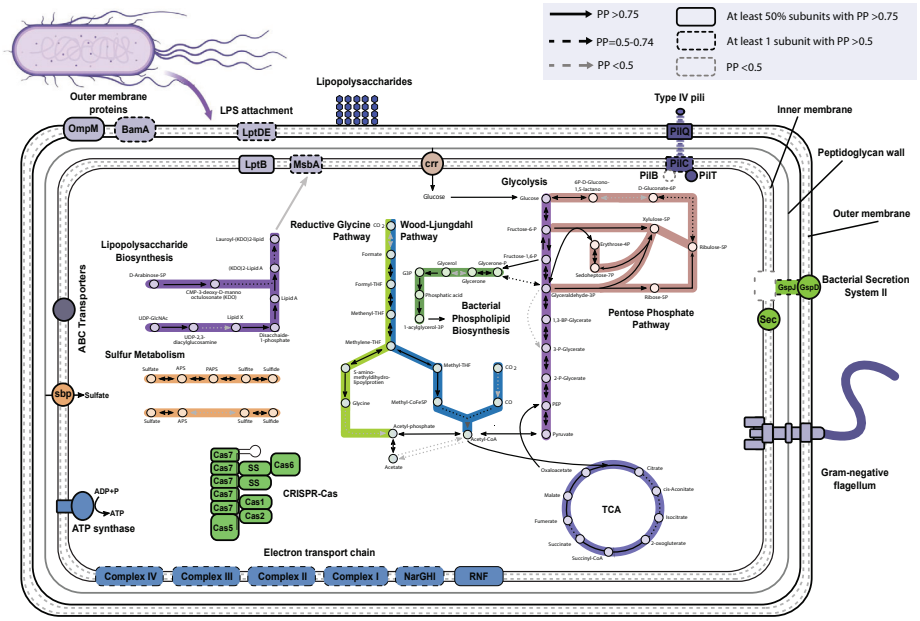


Fig. 4 | Ancestral reconstruction of the last bacterial common ancestor (LBCA). The reconstruction is based on genes that could be mapped to at least one branch within the root region with a PP > 0.5 (figs. S17 and S18) (*36*). The presence of a gene within a pathway is indicated as shown in the key. Our analyses suggest that LBCA was a rod-shaped, motile, flagellated double-membraned cell. We recover strong support for central carbon pathways, including glycolysis, the TCA cycle, and the pentose phosphate pathway. We did not find unequivocal evidence for the presence of a carbon fixation pathway, but we did find moderate support for components of both the WLP and the reverse TCA cycle. Although not depicted here, our analyses suggest that the machinery for core information processing and quorum sensing was also present in LBCA (table S7). As depicted in the key, arrows and boxes are shaded to indicate presence probabilities and, for multisubunit complexes, the number of subunits recovered with PP > 0.5. Note that the recovery of individual subunits for larger complexes does not imply that the complex was present [see online data supplement (*80*) for further discussion].

Moderate support for the presence of the shape-determining proteins MreB (PP = 0.9, 0.71, and 0.49 for roots 1 to 3, respectively, as depicted in Fig. 1B), MreC (PP = 0.82/0.78/ 0.68), and MreD (PP = 0.86/0.84/0.74) at the root suggests that LBCA was a rod-shaped cell (*52*). We also obtained high root PPs for proteins mediating outer cell envelope biosynthesis, including lipopolysaccharides (LPSs), from which we infer that LBCA had a double membrane with an LPS layer (*36*). Consistent with this inference, there was strong support for the flagellar subunits FlgH, FlgI, and FgA, which anchor flagella in diderm membranes (*54*), and for the type IV pilus subunit PilQ, which among extant bacteria is specific to diderms (*54*, *55*). Altogether, this supports hypotheses (*9*) in which LBCA was a diderm (*54–56*) and argues against scenarios in which the Gram-negative double membrane originated by endosymbiosis between monoderms [single-membraned bacteria (*10*)] or via the arrest of sporulation (*57*) in a spore-forming monoderm ancestor. Thus, diderm-to-monoderm transitions must have occurred subsequently on multiple occasions within Bacteria (*54–56*).

We recovered components of several core pathways for carbohydrate metabolism with high posterior support, including glycolysis, the tricarboxylic acid (TCA) cycle, and the pentose phosphate pathway (Fig. 4, figs. S17 and S18, and table S7) (*36*). Modern bacteria fix carbon using several different pathways, including the Calvin cycle, the 3-hydroxypropionate bicycle and variations thereof, the reductive glycine pathway (*58*), the Wood-Ljungdahl pathway (WLP), and the reverse TCA cycle, of which the latter two have been suggested to have emerged early in the history of life (*41*, *59*–*63*). Of these, we identified several enzymes of the TCA cycle and the reductive glycine pathway, although we did not recover the key enzymes of either pathway, and the directionality of the recovered enzymes is difficult to assess (*64*) (Fig. 4 and figs. S17 and S18). Furthermore, we identified several enzymes of the methyl branch of the WLP for acetate biosynthesis and components of a putative Rhodobacter nitrogen-fixing (RNF) complex (Fig. 4 and figs. S17 and S18), which together may indicate that LBCA was capable of acetogenic growth (*36*, *65*). However, the key enzyme of the WLP, the carbon monoxide dehydrogenase/acetyl coenzyme A synthase complex (*41*), had only moderate root support (PP = 0.5 to 0.75) for two subunits and low support (PP < 0.5) for other subunits. Thus, while our analyses support the antiquity of components of the WLP, acetogenesis, the TCA cycle, and several other core metabolic pathways, they do not confidently establish the combination of pathways used by LBCA (*36*).

Finally, our reconstruction also indicated high posterior support for elements of an adaptive immune CRISPR-Cas system (*66*, *67*), including the universally conserved Cas endonuclease, Cas1 (PP = 0.96/0.93/0.89), essential for spacer acquisition and insertion into CRISPR cassettes (*68*, *69*). Among other roles, CRISPR systems are crucial in antiviral defense and are activated in response to viral exposure (*70*); therefore, these findings are consistent with hypotheses suggesting that LBCA was already coevolving with parasitic replicators such as bacteriophages and plasmids (*71*, *72*).

VERTICAL AND HORIZONTAL EVOLUTION ARE COMPLEMENTARY

Here, we have used reconciliation methods to model both the vertical and horizontal components of bacterial evolution. These components are complementary, illuminating different facets of bacterial evolution, and we show that the horizontal component can be used to root and orient the vertical tree. Our analyses root the Bacteria between two major clades, the Terrabacteria and Gracilicutes, in contrast to recent outgroup-rooted analyses that place the root on the CPR branch. Instead, we predict that CPR evolved from a common ancestor with the Chloroflexota and Dormibacterota by reductive evolution. We infer that the last bacterial common ancestor was a fully fledged free-living diderm cell with an LPS layer, a multimeric flagellum, and a type III CRISPR-Cas system.

Phylogenetic models are necessarily simplified, and there is much work to be done to better capture the full heterogeneity of the evolutionary process in the reconciliation framework, from varying diversification rates to endosymbioses. With increased sampling and improved methods, reconciliation analyses should be able to probe still deeper into the early evolutionary history of life on Earth.

METHODS SUMMARY

PHYLOGENETICS

We used two alternative approaches to assemble representative sets of bacterial genomes. In the focal analysis, we sampled 265 genomes evenly from across the GTDB taxonomy (*13*). In the secondary analysis, we sampled 341 genomes according to the diversity of major bacterial lineages reported in a previous study (*11*). We used the OMA (*73*) algorithm to identify candidate single-copy orthologs and manually inspected initial single gene trees to identify a set of 62 congruent phylogenetic markers. Sequences were aligned using MAFFT 7.453 (*74*) and trimmed using BMGE 1.12 (*75*) with the BLOSUM30 matrix. Unrooted species trees were inferred from a concatenation of the 62 markers under the LG+C60+R8+F model in IQ-TREE 1.6.10 (*76*), which was the best-fitting model according to the BIC (*37*). To perform outgroup rooting analyses, we searched the genomes of 148 Archaea for orthologs of the 62-marker gene set and identified a subset of 29 genes with congruent single-gene phylogenies. AU tests (*28*) were performed in IQ-TREE.

GENE TREE–SPECIES TREE RECONCILIATION

To infer gene families, we performed all-versus-all DIAMOND (*77*) searches among the input protein sets and clustered the results using the MCL algorithm (*38*) with an inflation parameter of 1.2. Gene clusters were aligned and trimmed as described above, and bootstrap distributions inferred under the best-fitting model in IQ-TREE. We used ALEml_undated (*31*) to perform gene tree-species tree reconciliation. The relative ages of bacterial crown groups were estimated with MaxTiC (*43*) using only those transfer-based age constraints that were recovered in both the focal and secondary datasets. Estimates of gene family and lineage verticality were averaged over the reconciliations obtained in the focal analysis when rooting on each of the three candidate branches in the root region.

SIMULATIONS AND SENSITIVITY ANALYSES

To evaluate ALE performance, we simulated gene family evolution using Zombi (*78*) combined with rejection sampling to obtain sets of simulated gene families similar to the real data in terms of inferred DTL events. We then compared simulated and inferred numbers of events under a range of conditions (*36*). To evaluate the robustness of root inferences, we ordered gene families by decreasing DTL rates, rate ratios, and a range of other proxies for lack of informativeness and potential for introducing bias (*36*) and compared the likelihoods of competing root hypotheses as increasing proportions of gene families were excluded from the calculation.

ANCESTRAL METABOLIC RECONSTRUCTION

We inferred COG gene families by assigning the sequences on each sampled genome to gene families from the COG ontology (*39*) using eggNOG-mapper 2 (*79*), which were then used to perform gene tree-species tree reconciliation. Root origination probabilities for each of the 23 COG functional categories were inferred by maximizing the total reconciliation likelihood over all gene families in that category. These category-specific probabilities for origination

at the root were then used to estimate the PP that each gene family was present at the root of the tree. To infer the gene family content and metabolic repertoire of LBCA, functional annotations of protein sequences were obtained and assigned to COG families present at the root. The LBCA proteome was reconstructed taking into account the respective PPs for key gene families and metabolic pathways. A detailed account of all analyses is provided in the supplementary methods (*36*).

MATERIALS AND METHODS

TAXON SAMPLING FOR FOCAL ANALYSIS

To obtain a representative taxon sampling from across known bacterial diversity, we sampled taxa according to the classification provided by the Genome Taxonomy Database (GTDB r89) (*13*) as follows. First, we removed genomes with Quality < 0.75 (Quality is defined as Completeness - (5*Contamination) (*14*)), and filtered out all phyla subsequently left with fewer than 10 species. Genomes were sampled from the remaining taxa on a per-class basis: for classes containing a single order, the genome with the highest quality score was sampled; for classes containing multiple orders, the highest quality genome from each of two randomly chosen orders was sampled. This protocol ensured that every class in the GTDB is represented in the final tree. We then manually added the genome of *Gloeomargarita litophora* given its importance in constraining the phylogeny and timing of chloroplast evolution. The list of genomes can be found in table S9.

UNROOTED SPECIES TREE INFERENCE

We used Orthologous Matrix (OMA) 2.1.1 (*73*) to identify candidate single- copy bacterial orthologs, and retained those with at least 75% of all species represented in each family. Sequences were aligned in MAFFT *(74)* using the -auto option, and trimmed in BMGE 1.12 (*75*) using the BLOSUM30 model. Initial trees were inferred for each candidate marker gene under the LG+G+F model in IQ-TREE 1.6.10 *(76)*. The trees were manually inspected, and we selected orthologues where the monophyly of 14 pre-defined major lineages was not violated with bootstrap support >70%, resulting in 62 final orthologues. Concatenation of this marker set resulted in an alignment of 18,234 amino acids. We inferred an unrooted phylogeny from this concatenate under the LG+C60+R8+F model, which was chosen as the best-fitting model by the BIC criterion in IQ-TREE (*76*). We additionally removed the most compositionally heterogeneous sites from the sequence alignment using Alignment Pruner (*81*) (https://github.com/novigit/davinciCode/blob/master/perl) (20%, 40%, 60% and 80% respectively) and inferred trees using the same procedure described above in order to compare the resulting topologies.

OUTGROUP ROOTING

To root the bacterial tree using an archaeal outgroup, we used a representative sampling of 148 archaeal genomes and inferred the ML tree in IQ-TREE under the best-fitting LG+C60+R8+F model. The concatenated alignment included a subset of 29 out of the 62 bacterial orthologs

that were shared between bacteria and archaea, as determined by by Hidden Markov Model (HMM) searches and manual inspection of single gene trees. ML trees for these 29 genes recovered the clanhood (*82*) of Bacteria and Archaea. We performed approximately-unbiased (AU) tests *(28)* to determine whether a range of published alternative rooting hypotheses (table S2) could be rejected, given the model and data (AU p-value > 0.05).

GENE FAMILY CLUSTERING AND ALE ANALYSIS

We used the protein annotations provided by GTDB, which were originally obtained using Prodigal. To infer homologous gene families for amalgamated likelihood estimation (ALE), we performed an all vs all similarity search using Diamond (*77*) with an E-value threshold of $<10^{-7}$ to avoid distant hits and k = 0 to report all the relevant hits. Current clustering methods are not consummate and the parameters that determine the granularity of clustering do not have a direct biological motivation. Setting the value of the Markov Cluster (MCL) algorithm *(83)* inflation parameter therefore involves a trade-off between inferring large, inclusive clusters that will contain false positives (sequences that are not part of the real gene family) and small, conservative clusters that may divide real gene families into several subclusters. An additional practical concern for phylogenomics is that overly large clusters may align poorly and result in low-quality single protein trees. In our rooting analysis, we experimented with a range of values for the mcl inflation parameter, and chose 1.2 because the clusters were inclusive without a substantial reduction in post-masking alignment length compared to more granular settings.

Clustering using MCL (*83*) with an inflation parameter of 1.2 resulted in 186,827 gene families and a total of 11,765 families with 4 or more sequences. We aligned the 11,765 gene families using MAFFT (*74*) (with the --auto option) and filtered with BMGE (*75*) (using bmge -t AA -m BLOSUM30) After filtering, 260 alignments contained no high-quality columns and were discarded. We filtered out sequences comprising more than 80% gaps to produce the final set of alignments. We also discarded all alignments with less than 30 columns, leaving a total of 11,272 families. The gene trees were computed using IQ-TREE v 1.6.10 using the following command: iqtree -m TEST -s FAMXXX.faa.aln.trimmed -bb 10000 -wbtl -nt AUTO -madd LG4X,LG4M,LG+C10,LG+C20,LG+C30,LG+C40,LG+C50,LG+C60,C10,C20,C30,C40,C50,C60.

Conditional clade probabilities (CCPs) were computed using ALEobserve and the resulting ALE files were reconciled with the species tree. Loss rates were corrected by genome completeness, estimated using CheckM (*84*). We tested 62 roots (Online Data Supplement *(80)*).

SIMULATIONS TO EVALUATE THE PERFORMANCE OF ALE

Performance of ALE for species tree rooting

The ability of the ALEml_undated algorithm to infer the correct gene tree root in the presence of gene duplications, transfers and losses was previously investigated using simulations (*1*). Briefly, gene families were simulated on a rooted species tree using a continuous-time origination, duplication, transfer and loss (ODTL) process (that is, a more complex model of

genome evolution than that implemented in ALEml_undated), and ALEml_undated was used to estimate the root from subsamples of the simulated families. The maximum likelihood root according to ALE was the correct root in 95/100 replicates, and the log likelihood of alternative roots decreased with nodal distance from the correct root (as observed in our empirical data, see Fig. 1). In the remaining 5 cases, the maximum likelihood root was one branch away from the true root. Analysis of empirical data suggested that ALE root inferences are robust to (that is, consistent across) subsets of the data that vary in terms of the rate of horizontal gene transfer or species representation in gene families (*1*). These properties make the ALE approach appropriate for inferring the root of Bacteria.

Accuracy of duplication, transfer and loss rates inferred by ALE

To test the accuracy of ALE at correctly inferring duplication, transfer and losses, we simulated a species tree of 265 leaves (the same size as the focal dataset) using Zombi (*78*). A simulation approach is necessary because, for empirical data, we do not know the true gene family history and so cannot evaluate method performance directly. The empirical realism of simulations is often an issue, and it is not always clear how best to accommodate the complexities of real genome evolution, including heterogeneity of DTL rates across families and, indeed, biases in rates (for instance, the high lineage-specific rate of gene loss that appears to characterise CPR). To make our simulations as realistic as possible, we first simulated the evolution of gene families using the Gm mode in Zombi, which assumes that every family has its own and independent rates of D, T and L, with the rates sampled from distinct gamma distributions (D~G(0.2,0.5), T~G(2,05), L~G(2.2, 0.5)). We simulated a total of 97929 families. We computed gene tree-species tree reconciliations with ALEml_undated for all families. Then, to ensure that the simulated dataset was as similar as possible to the real dataset, we sampled 2000 families at random from the real dataset. For each of those families, we selected the simulated family most similar in terms of DTL events (similarity was computed as one minus the squared sum of differences between the different inferred events by the reconciliations and the size of the families). This procedure resulted in a set of simulated families that were closely similar to the real data in terms of DTL events (fig. S3). These families recapitulate the gene family- and lineage-specific heterogeneity of DTL rates observed in the real data, and so provide the best possible basis for evaluating the performance of ALE. We used this set of simulated families in the analyses described below.

Comparison of the real (that is, simulated) and inferred numbers of D, T and L events on these data suggest that ALE accurately estimates the numbers of all three kinds of events (fig. S4), with mean errors close to 0 for all three types of events (D ~ 0.005, T ~ 0.048, L ~ 0.019). A detailed examination of the errors that do occur indicated that errors are most common in small families (fig. S5), and that the number of DTL events tends to be under-estimated when the true number of events is high (fig. S6). These observations motivate some of the sensitivity analyses of the empirical data described below, in which gene families with high inferred rates and, in a separate analysis, small sizes were excluded from the root calculation (see “Testing

the robustness of the inferred root region" below); the Gracilicutes-Terrabacteria root was robust to all of these treatments.

Different combinations of DTL events can give rise to the same gene tree topology. For example, genes that are patchily distributed across species might be explained by a series of gene transfers, ancestral presence followed by independent losses, or a combination of processes. To investigate whether ALE can distinguish between different kinds of DTL events based on gene tree topologies, we examined the correlations in inference errors for different kinds of events. Negative correlations (for example, over-estimation of transfer associated with under-estimation of losses) would suggest that the method can mistake one kind of DTL event for another. No correlations of this type were obtained (fig. S7), suggesting that ALE can distinguish the history of DTL events giving rise to a given gene tree topology. To further investigate whether ALE overestimates the number of gene transfers compared to duplications and losses, we specifically examined the inference results for the subset of simulated families with 0 transfers but one or more duplication and loss events (2429 of 97292 families). Of these 2,429 families, ALE correctly inferred that 2,332 (96%) had no transfers.

Next, we investigated whether biases in DTL rates across the tree - which result in variation in genome sizes, such as the small genomes of CPR - impact ALE inference accuracy. To do so, we performed two additional simulations: one in which gene family originations occur at random on the tree (which results in homogeneous simulated gene contents and genome sizes), and one in which gene originations were constrained to occur at the same points as they do in the real data. This latter simulation results in data that recapitulate the variation in gene content and genome size observed in the empirical data. We then compared inference accuracy on the two datasets (Fig. S8). The results suggest that genome size heterogeneity does not substantially affect inference accuracy, with all errors centred on 0 in both datasets (mean errors without heterogeneity: D ~ 0.00342, T ~ 0.0044, L ~ 0.0238; with heterogeneity: D ~ 0.00349, T ~ -0.0007, L ~ -0.0268).

Finally, we investigated the impact of lineage extinction on the accuracy of DTL estimates. In principle, ALE estimates ought to be robust to lineage extinction because gene acquisitions from extinct (or unsampled) lineages are accounted for in the method (*85*). To investigate, we used Zombi (*78*) to simulate 1000 species trees with 30 extant (sampled) taxa, with the speciation rate equal to the extinction rate. We performed simulations on species trees with relatively small numbers of tips because, since Zombi is a forward simulator, simulating trees with 265-341 tips in the context of high extinction rates is computationally intractable (that is, only a very small proportion of simulated trees will grow to have 100s of extant tips when the extinction rate is equal to or larger than the speciation rate). 100 gene families were simulated on each species tree; of these, 69921 had at least 4 surviving gene copies in the extant tip genomes and could be used for comparison of DTL inference accuracy. For each family, we calculated inference accuracy as (inferred number of events - simulated number of events)/family size, as above, and evaluated the relationship between the proportion of

extinct lineages on the species tree and inference accuracy (fig. S2). We detected a statistically significant but quantitatively small impact of extinction on accuracy, with a higher proportion of extinct lineages corresponding to a slight increase in error (Correlation coefficients between accuracy and proportion of extinct lineages: -0.024 (D); 0.041 (T); -0.093 (L); fig. S2); errors remain centred on 0 even when almost all lineages had gone extinct.

TESTING THE ROBUSTNESS OF THE INFERRED ROOT REGION

Simulations (see above) are the most direct way to evaluate the performance of ALE, because they provide a controlled situation in which we know the truth with certainty. However, real data are heterogeneous in ways that are difficult to recapitulate in simulations, particularly in terms of variation and biases in the rates of evolutionary processes (DTL, speciation, extinction and substitution rates) across the tree. We therefore performed a range of sensitivity analyses to evaluate the robustness of the inferred root region.

Distributions of DTL rates, rate ratios, and the impact of excluding gene families from the root calculation based on these and other criteria.

Firstly, we ranked gene families by inferred duplication, transfer and loss rates and rate ratios (figs S10-11), and performed a gene-filtering analysis (fig. S12) in which families at either end of the distribution were progressively removed and root likelihoods re-evaluated. This approach is analogous to fast site removal, in that families with very high or low rates may be difficult to model and so mislead inference.

Based on these rankings, we performed gene filtering analyses in which the highest ranked families were progressively removed and the difference in likelihoods between roots (ΔLL) re-evaluated (fig. S12A-S). The first of these plots (fig. S12B-C) illustrates the effect of progressively filtering out the most widely-distributed families (the metric is the number of species with at least one gene in the family). For each pair of plots, the bottom panel shows the threshold value corresponding to the percentile removed. For example, removing 5% of the families represented in the most species corresponds to a threshold of being represented in 150 species or more for MCL families. Note that these threshold plots can also be interpreted as the cumulative distribution of the ranking criteria, i.e. the above example implies that 5% of MCL families are represented on 150 or more genomes. The left hand side of each plot indicates the summed likelihood of each candidate root position (Fig. S12A) on all of the data; moving to the right along the x-axis illustrates how the summed likelihood for each root changes as families at the top end of the distribution are filtered out. Filtering out broadly-distributed families has a similar effect in both the MCL and Clusters of Orthologous Genes (COG) datasets: ΔLL starts to diminish. That is, ALE starts to lose the ability to distinguish between different root hypotheses. Conversely, when ranking families by the opposite criterion (the number of species absent, fig. S12D-E), the difference in likelihood between the roots remains unchanged until a very large fraction of families is excluded.

We next evaluated whether families with different estimated D, T or L rates agree on the optimal root region (that is, whether the root signal from families with high, moderate and low rates is consistent). To evaluate the signal in a root-independent manner, we calculated the mean per family D, T and L rates weighted according to the likelihood of that family for different roots. We can see in fig 12F-K that removing even a substantial (10-20%) fraction of families with the highest D, T or L rates does not change the order of likelihoods for different roots, and it is only after nearly half the families are discarded that we start to see a loss of resolution.

To filter families with potentially problematic rates, we also calculated each of the 6 possible rate ratios (D/T, T/D, D/L, L/D, T/L and L/T) and ranked families according to the maximum of these six ratios (fig. S12L-M). While our results remain unchanged as long as fewer than 10% of families with the most extreme rate ratios are removed, removing between 10 and 30% of the families with the most extreme rate ratios changed the order of the most likely roots while still retaining resolution (see for colors). This indicates that the Fusobacteriota root (dark green) derives its support, in part, from families that are outliers in terms of DTL rate ratios.

Performing further analogous threshold analyses (fig. S12N-S) in terms of mean copy number and verticality, we find that the rooting analysis is robust to filtering based on these criteria.

Effect of gene family clustering

We next evaluated the impact of gene family clustering on our analysis by repeating the entire analysis using COG (*39*) families; the same root region was recovered with the addition of one adjacent branch, with the reduced resolution likely due to the smaller size of the COG family dataset (3,723 vs 11,272 mcl families; see table S5).

Effect of taxon sampling

Taxon sampling is known to be an important factor in phylogenetic analyses (*22*), so we next evaluated whether our method of sampling taxa representatively from GTDB impacted our results. To do so, we repeated the entire analysis using a different approach, selecting representative taxa from major bacterial clades previously described in the literature *(11,15)*. We based our sampling on the analysis of (*11*), in which CPR comprise 40% of bacterial diversity. To do so, we inferred a tree of the bacterial portion of the published (*11*) concatenate under the LG+G4+F model in IQ-TREE. We divided the tree into 7 major bacterial clades based on a literature search (table S12) and additional environmental lineages with branch length diversity comparable to the known groups. For each group defined in this way, we manually subsampled taxa so as to maintain genetic diversity, while avoiding the longest and shortest branches. We sampled 341 species, comprising 200 'classic' bacteria, 124 CPR bacteria and one bacterial genome respectively from each of the 17 new phyla described by (*14*); see table S9. We used the same marker gene set as in the focal analysis. A species tree was inferred in IQ-TREE using the LG+C20+G4 model with PMSF *(86)*. Additional trees were inferred in PhyloBayes under the CAT+GTR+G4 model using a recoded alignment using the four-category scheme of Susko and Roger *(87)*, and under the multispecies coalescent model in ASTRAL (*88*). To

infer homologous gene families, we used the same pipeline as that used in the focal analysis. This resulted in 11,781 gene families with 4 or more sequences, which were analysed as in the focal analysis. The unrooted species trees are congruent with each other, except in the placement of a small group of phyla comprising Fusobacteriota, Aquificota, Synergistota, Spirochaetota and Thermtogota ("FASST"). These taxa are resolved in different positions in each of the three unrooted topologies, and are found to be monophyletic in the tree inferred fromthe recoded alignment. Similarly to the focal analysis, the ALE analyses yielded two root positions that could not be rejected (AU test, $p > 0.05$), summarised in fig. S13. Both of these rooted phylogenies are congruent with that of the focal (GTDB) analysis, with Terrabacteria and Gracilicutes on either side of the root, with the only differences being in the placement of the FASST taxa. In the focal analysis, as well as the LG+PMSF+G4 and ASTRAL trees inferred as part of the GTDB-independent analysis, FASST were not recovered as a monophyletic group.

INFERENCE OF RELATIVE DIVERGENCE TIMES OF BACTERIAL CLADES

We applied the pipeline documented and implemented in the Online Data Supplement (*80*, RelativeDating.zip) to obtain the relative dated trees. We performed the analysis 5 times: 3 times in the focal dataset, and 2 times in the secondary dataset (using all the roots that could not be rejected). The pipeline consists of parsing the transfers inferred by ALEml_undated (using the MCL families) and discarding those with posterior probability < 0.05. We used bootstrapping to estimate constraint support in the following way: for each of the three branches in the root region, we sampled the gene families 100 times with replacement and, for each replicate, converted detected transfers to constraints and performed a Maximum Time Consistency (MaxTiC) analysis (*42*, *43)*. We then selected the phyla that were represented by 5 genomes or more in both datasets. We pruned the tree of the focal dataset to maintain only those phyla. Finally, we selected the constraints that were supported by both datasets and use those constraints to generate 1000 time orders compatible with those constraints the script order_explorer.py *(80)*). We then ranked all interior nodes on the tree, with the root node having rank 0 and the most recent speciation node having rank 11.

QUANTIFYING VERTICAL AND HORIZONTAL SIGNALS IN BACTERIAL GENOME EVOLUTION

In the context of our analyses, "verticality" is the proportion of inferred evolutionary events on a branch of the rooted species tree that reflect vertical descent, estimated using gene tree-species tree reconciliation. We defined branch-wise verticality as V/(V+O+T), where V is the inferred number of vertical transmissions of a gene from the ancestral to descendant ends of the branch; O is the number of new gene originations on the branch; and T is the number of gene transfers into the branch. We defined transfer propensity as T/(V+T), where V and T refer to inferred numbers of events within the history of a gene family (table S11).The numbers reported in the main text have been averaged over the reconciliations obtained using the three possible roots of the focal analysis.

ANCESTRAL GENE CONTENT AND METABOLIC RECONSTRUCTION

Protein and protein family functional annotation

Protein sequences from all genomes used for phylogenetic analyses in this study were annotated using a variety of databases. Functional annotations were obtained using hmmsearch v3.1b2 (settings: -E 1e-5) (*89,90*) against KEGG Orthology (KO) annotations from the KEGG Automatic Annotation Server (KAAS; downloaded April 2019) (*91*). Additionally, all proteins were scanned for protein domains using InterProScan (v5.31-70.0; settings: --iprlookup --goterms) (*92*).

Multiple hits corresponding to the individual domains of a protein are reported using a custom script (parse_IPRdomains_vs2_GO_2.py). For the functional annotation of the 4256 COG families investigated in our ancestral reconstructions, we assigned Kos using a majority rule, i.e. we assigned the KO reported in the majority of sequences comprising each of the COG families yielding a COG-to-KO mapping file. Subsequently, we mapped COG descriptions, COG Process/Class, Category description, kegg id, kegg description, and kegg pathway to the COG-to-KO mapping file. COG descriptions were collected from the root annotations (1_annotations.tsv) downloaded at EggNOG (v5.0.0) (*93*). COG functional category and Process/Class descriptions were derived from eggNOG (v4.0) (*94*). KO pathways were manually curated based on an in-house KO-to-pathway mapping file, and were subsequently mapped to the respective KO. The scripts for annotation and mapping are included in the Online Data Supplement *(80)*.

COG gene families for ancestral gene content reconstruction

We built a set of gene families based on the COG (*95*) database for ancestral functional inference. To do so, we annotated each genome in the dataset using eggNOG-mapper v2(*79*), then clustered proteins into families based on their COG annotations. For proteins annotated with more than one COG category (8% of proteins), we included the protein in both COG families. This resulted in 4256 COG families, of which 3723 had 4 or more sequences. COG families are ideal for ancestral reconstruction because they comprise all of the sequences on extant genomes that can be annotated with a given unambiguous function from the COG ontology. In addition, the hierarchical nature of the COG classification (comprising gene family annotations nested within 23 broader functional categories) enabled us to explicitly model the different evolutionary ages of gene functional classes as part of the analysis, by using category-specific root origination priors (see below).

Our COG families are useful for functional reconstruction, but are perhaps less well suited for investigating other aspects of bacterial evolution because they are constructed only from proteins that could be annotated with eggnog mapper. By contrast, MCL families represent --- within the limitations of the clustering approach, as discussed above --- an unbiased view of gene family diversity for the set of genomes we analyzed. We therefore base analyses other than those regarding the functional annotation of LBCA on the MCL families. However, since

gene clustering methods are not consummate and each has strengths and weaknesses, we also investigated the root signal from the COG families. This analysis identified a similar root region of four adjacent branches, comprising the root region from the focal analysis (3 branches) plus one additional root, in which Spirochaetota branched on the Terrabacteria side of the root (table S5).

Root gene mapping approach

To estimate root presence posterior probabilities (PPs) for each gene family for each of the three supported roots, we first estimated the probability of origination at the root (O_R) by maximum likelihood, finding the O_R value that maximises the total reconciliation likelihood summed over all gene families (table S11). We then used the global ML O_R value to calculate the root presence posterior probabilities for each family; that is, the probability that one or more copies of a given gene family were present at the root, given the ML O_R value. We estimated root origination rates independently for each of the 23 COG functional categories, and used these rates to estimate the posterior probability of presence at the root node for each gene family. Note that, for all nodes of the tree (including the root nodes), we additionally estimated PPs directly from the sampled reconciliations. Python code implementing this procedure is provided in (*80*) at Code/O_R_Optimization.py.

Initial gene content and metabolic inferences at a particular node were based on gene families with a posterior presence probability (PP) of >0.95 at that node. This approach is conservative and could miss the presence of certain pathways which may be represented by proteins with a range of PP values. Therefore, we manually investigated the PPs of pathways discussed in this manuscript and inferred the presence of specific pathways or functional modules if the majority of the components were found with PP >0.50, as described in the main text, Fig. 4 and fig. S17.

Impact of root branch on LBCA gene content

The credible set of root branches from the ALE analysis comprised three adjacent branches at the centre of the tree (Fig. 1b). The difference between these three root positions relates to the placement of Fusobacteria, either as the root branch or as the most basal split on either the Gracilicutes or Terrabacteria+DST “sides” of the rooted tree. We therefore estimated root PPs for COG families on all three branches; root PPs under all three roots are provided in the Online Data Supplement *(80)*.

Metabolic comparisons

Results from the PP analysis were used as the framework for metabolic comparisons and reconstruction of the proteome of LBCA. First, the occurrence of an individual COG family across each taxon was counted in R (v3.6.3) (table S4). This binary presence/absence matrix was combined with the PP values for Nodes corresponding to the CPR, Chloroflexota+CPR, Chloroflexota, Terrabacteria, DST+Terrabacteria, Gracilicutes-Spirochaetota, Gracilicutes+Spirochaetota, Root 1, Root 2, and Root 3, filtered with a cutoff of PP>0.50. The

combined count table was summarized using the ddply function of the plyr package (v1.8.4), which was used to summarize the counts across each phylogenetic cluster, node, and root. Data is visualized in a heatmap generated using the ggplot function with geom_tile and facet_grid of the ggplot2 package (v3.2.0). Heatmap categories for pathways were scaled based on the number of COG families, results were plotted using the grid.draw function of the grid package (v3.6.3). Heatmaps were manually merged with a species tree in Adobe Illustrator (v22.0.1).

Evaluating the robustness of the LBCA reconstruction to taxon sampling based on the secondary dataset

To evaluate the impact of taxon sampling, we analyzed both the primary and secondary datasets with the same ancestral reconstruction pipeline. Across all protein families, agreement between the datasets was highly significant, though lower than among the root regions for each dataset (table S13). The PPs between the focal and secondary analysis were significantly correlated independent of rooting (Pearson's correlation 0.48-0.53 $p<10^{-16}$, see table S13). PPs between root positions in the context of the same analysis were very strongly correlated, with the highest correlation between root 1 and root 2 of the secondary analysis (0.96) and roots 1 and 2 of the primary analysis 0.94). Overall, root 3 of the focal analysis (corresponding to the Fusobacteriota root) correlated the least with other roots. The focal analysis compared to the secondary analysis considered fewer taxa (265 vs 341) and correspondingly fewer COGs (3782 vs 4220 with four or more genes). The focal analysis also recovered fewer COGs with high confidence (PP>0.9) at the root (see gray diagonal fields in table S13). Of the COGs recovered with high confidence at the root in the focal analysis, the majority (63%-78%) were also recovered with high confidence at the root in the secondary analysis (see gray off-diagonal fields in table S13). Considering COGs recovered with high confidence at the root, root 3 of the focal analysis is again the least congruent with the other gene sets, and at the same time is the one with the smallest number of COGs recovered at the root. Importantly, all of the ancestral pathways discussed in the main text were recovered at the root with moderate to high PP support in both datasets, as indicated (table S7).

ACKNOWLEDGMENTS

FUNDING

G.A.C. is supported by a Royal Society Research Grant to T.A.W. T.A.W. is supported by a Royal Society University Research Fellowship and NERC grant NE/P00251X/1. G.J.Sz. received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program under grant agreement 714774 and grant GINOP-2.3.2.-15-2016-00057. A.S. is supported by the Swedish Research Council (VR starting grant 2016-03559 to A.S.) and the NWO-I foundation of the Netherlands Organisation for Scientific Research (WISE fellowship to A.S.). A.A.D. and P.H. are supported by an Australian Research Council Laureate Fellowship (grant FL150100038).

AUTHOR CONTRIBUTIONS

The project was conceived of by T.A.W., G.J.Sz., P.H., A.S., G.A.C., and A.A.D. G.A.C., A.A.D., T.A.W., L.L.Sz., and G.J.Sz. performed phylogenomic analyses. G.J.Sz. developed new analytical methods. T.A.M., G.A.C., A.A.D., and A.S. performed metabolic annotations and reconstructions. All authors contributed to interpretation and writing.

COMPETING INTERESTS

The authors have no competing interests to declare.

DATA AND MATERIALS AVAILABILITY

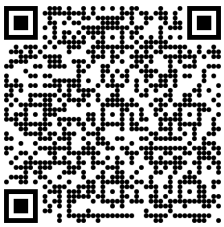
All data and code are provided online at Figshare (*80*). New methods are described in detail in the supplementary methods (*36*).

SUPPLEMENTARY MATERIAL

SUMMARY

Materials and Methods
Figs. S1 to S18
Tables S1 to S13
References (*81–95*)
MDAR Reproducibility Checklist

All supplementary information and files be accessed here:
https://www.science.org/doi/10.1126/science.abe0511#supplementary-materials



Correction (30 June 2021): In Fig. 4, “OmpM” should read “OmpH,” and several metabolic interconversions were shaded incorrectly. A sentence has been added to the Fig. 4 legend to better explain the key. The PP values in one sentence on page 6 were partially incorrect; the sentence has been corrected to read ““MreB (PP = 0.9, 0.71, and 0.49 for roots 1 to 3, respectively, as depicted in Fig. 1B), MreC (PP = 0.82/0.78/ 0.68), and MreD (PP = 0.86/0.84/0.74).” The underlying data (table S7) are correct as originally published. Figure 4, fig. S17 (on which Fig. 4 was partially based), and the aforementioned sentence have been corrected in both the PDF and HTML versions online.

REFERENCES

1. T. A. Williams, G. J. Szöllősi, A. Spang, P. G. Foster, S. E. Heaps, B. Boussau, T. J. G. Ettema, T. M. Embley, Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E4602–E4611 (2017).

2. J. R. Brown, W. F. Doolittle, Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 2441–2445 (1995).

3. J. P. Gogarten, H. Kibak, P. Dittrich, L. Taiz, E. J. Bowman, B. J. Bowman, M. F. Manolson, R. J. Poole, T. Date, T. Oshima, J. Konishi, K. Denda, M. Yoshida, Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 6661–6665 (1989).

4. N. Iwabe, K. Kuma, M. Hasegawa, S. Osawa, T. Miyata, Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 9355–9359 (1989).

5. O. Zhaxybayeva, P. Lapierre, J. P. Gogarten, Ancient gene duplications and the root(s) of the tree of life. *Protoplasma* **227**, 53–64 (2005).

6. F. U. Battistuzzi, S. B. Hedges, A major clade of prokaryotes with ancient adaptations to life on land. *Mol. Biol. Evol.* **26**, 335–343 (2009).

7. M. Bocchetta, S. Gribaldo, A. Sanangelantoni, P. Cammarano, Phylogenetic depth of the bacterial genera Aquifex and Thermotoga inferred from analysis of ribosomal protein, elongation factor, and RNA polymerase subunit sequences. *J. Mol. Evol.* **50**, 366–380 (2000).

8. C. Brochier, H. Philippe, Phylogeny: a non-hyperthermophilic ancestor for bacteria. *Nature* **417**, 244 (2002).

9. T. Cavalier-Smith, Rooting the tree of life by transition analyses. *Biol. Direct* **1**, 19 (2006).

10. J. A. Lake, Evidence for an early prokaryotic endosymbiosis. *Nature* **460**, 967–971 (2009).

11. L. A. Hug, B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D. A. Relman, K. M. Finstad, R. Amundson, B. C. Thomas, J. F. Banfield, A new view of the tree of life. *Nat Microbiol* **1**, 16048 (2016).

12. S. Mukherjee, R. Seshadri, N. J. Varghese, E. A. Eloe-Fadrosh, J. P. Meier-Kolthoff, M. Göker, R. C. Coates, M. Hadjithomas, G. A. Pavlopoulos, D. Paez-Espino, Y. Yoshikuni, A. Visel, W. B. Whitman, G. M. Garrity, J. A. Eisen, P. Hugenholtz, A. Pati, N. N. Ivanova, T. Woyke, H.-P. Klenk, N. C. Kyrpides, 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat. Biotechnol.* **35**, 676–683 (2017).

13. D. H. Parks, M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil, P. Hugenholtz, A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).

14. D. H. Parks, C. Rinke, M. Chuvochina, P.-A. Chaumeil, B. J. Woodcroft, P. N. Evans, P. Hugenholtz, G. W. Tyson, Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* **2**, 1533–1542 (2017).

15. C. T. Brown, L. A. Hug, B. C. Thomas, I. Sharon, C. J. Castelle, A. Singh, M. J. Wilkins, K. C. Wrighton, K. H. Williams, J. F. Banfield, Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).

16. Q. Zhu, U. Mai, W. Pfeiffer, S. Janssen, F. Asnicar, J. G. Sanders, P. Belda-Ferre, G. A. Al-Ghalith, E. Kopylova, D. McDonald, T. Kosciolek, J. B. Yin, S. Huang, N. Salam, J.-Y. Jiao, Z. Wu, Z. Z. Xu, K. Cantrell, Y. Yang, E. Sayyari, M. Rabiee, J. T. Morton, S. Podell, D. Knights, W.-J. Li, C. Huttenhower, N. Segata, L. Smarr, S. Mirarab, R. Knight, Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477 (2019).

17. C. J. Castelle, J. F. Banfield, Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. *Cell* **172**, 1181–1197 (2018).

18. C. J. Castelle, C. T. Brown, K. Anantharaman, A. J. Probst, R. H. Huang, J. F. Banfield, Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat. Rev. Microbiol.* **16**, 629–645 (2018).

19. J. P. Beam, E. D. Becraft, J. M. Brown, F. Schulz, J. K. Jarett, O. Bezuidt, N. J. Poulton, K. Clark, P. F. Dunfield, N. V. Ravin, J. R. Spear, B. P. Hedlund, K. A. Kormas, S. M. Sievert, M. S. Elshahed, H. A. Barton, M. B. Stott, J. A. Eisen, D. P. Moser, T. C. Onstott, T. Woyke, R. Stepanauskas, Ancestral absence of electron transport chains in Patescibacteria and DPANN. *Front. Microbiol.* **11**, 1848 (2020).

20. C. J. Castelle, K. C. Wrighton, B. C. Thomas, L. A. Hug, C. T. Brown, M. J. Wilkins, K. R. Frischkorn, S. G. Tringe, A. Singh, L. M. Markillie, R. C. Taylor, K. H. Williams, J. F. Banfield, Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr. Biol.* **25**, 690–701 (2015).

21. R. Méheust, D. Burstein, C. J. Castelle, J. F. Banfield, The distinction of CPR bacteria from other bacteria based on protein family content. *Nat. Commun.* **10**, 4173 (2019).

22. A. Graybeal, Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst. Biol.* **47**, 9–17 (1998).

23. S. M. Hedtke, T. M. Townsend, D. M. Hillis, Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst. Biol.* **55**, 522–529 (2006).

24. J. Bergsten, A review of long-branch attraction. *Cladistics* **21**, 163–193 (2005).

25. R. Gouy, D. Baurain, H. Philippe, Rooting the tree of life: the phylogenetic jury is still out. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140329 (2015).

26. J. A. Lake, R. G. Skophammer, C. W. Herbold, J. A. Servin, Genome beginnings: rooting the tree of life. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 2177–2185 (2009).

27. R. G. Skophammer, J. A. Servin, C. W. Herbold, J. A. Lake, Evidence for a gram-positive, eubacterial root of the tree of life. *Mol. Biol. Evol.* **24**, 1761–1768 (2007).

28. H. Shimodaira, An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508 (2002).

29. G. J. Szöllosi, B. Boussau, S. S. Abby, E. Tannier, V. Daubin, Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17513–17518 (2012).

30. L. A. David, E. J. Alm, Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* **469**, 93–96 (2011).

31. G. J. Szöllõsi, W. Rosikiewicz, B. Boussau, E. Tannier, V. Daubin, Efficient exploration of the space of reconciled gene trees. *Syst. Biol.* **62**, 901–912 (2013).

32. L. A. Katz, J. R. Grant, L. W. Parfrey, J. G. Burleigh, Turning the crown upside down: gene tree parsimony roots the eukaryotic tree of life. *Syst. Biol.* **61**, 653–660 (2012).

33. D. M. Emms, S. Kelly, STRIDE: Species tree root inference from gene duplication events. *Mol. Biol. Evol.* **34**, 3267–3278 (2017).

34. A. Zwaenepoel, Y. Van de Peer, Inference of ancient whole-genome duplications and the evolution of gene duplication and loss rates. *Mol. Biol. Evol.* **36**, 1384–1404 (2019).

35. S. Q. Le, O. Gascuel, An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).

36. G. A. Coleman, A. A. Davín, T. A. Mahendrarajah, L. L. Szánthó, A. Spang, P. Hugenholtz, G. J. Szöllősi, T. A. Williams, Detailed Materials and Methods. (2021). https://www.science.org/doi/10.1126/science.abe0511#supplementary-materials

37. D. Posada, T. R. Buckley, Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst. Biol.* **53**, 793–808 (2004).

38. A. J. Enright, S. Van Dongen, C. A. Ouzounis, An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).

39. M. Y. Galperin, D. M. Kristensen, K. S. Makarova, Y. I. Wolf, E. V. Koonin, Microbial genome analysis: the COG approach. *Brief. Bioinform.* **20**, 1063–1070 (2019).

40. K. Raymann, C. Brochier-Armanet, S. Gribaldo, The two-domain tree of life is linked to a new root for the Archaea. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 6670–6675 (2015).

41. P. S. Adam, G. Borrel, S. Gribaldo, Evolutionary history of carbon monoxide dehydrogenase/acetyl-CoA synthase, one of the oldest enzymatic complexes. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E1166–E1173 (2018).

42. A. A. Davín, E. Tannier, T. A. Williams, B. Boussau, V. Daubin, G. J. Szöllősi, Gene transfers can date the tree of life. *Nat Ecol Evol* **2**, 904–909 (2018).

43. C. Chauve, A. Rafiey, A. A. Davín, C. Scornavacca, P. Veber, B. Boussau, G. J. Szöllősi, V. Daubin, E. Tannier, MaxTiC: Fast ranking of a phylogenetic tree by Maximum Time Consistency with lateral gene transfers, *bioRxiv* (2017). https://doi.org/10.1101/127548.

44. T. Dagan, W. Martin, Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 870–875 (2007).

45. W. F. Doolittle, Phylogenetic classification and the universal tree. *Science* **284**, 2124–2129 (1999).

46. W. F. Doolittle, E. Bapteste, Pattern pluralism and the Tree of Life hypothesis. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 2043–2049 (2007).

47. T. Dagan, W. Martin, The tree of one percent. *Genome Biol.* **7**, 118 (2006).

48. D. Alvarez-Ponce, P. Lopez, E. Bapteste, J. O. McInerney, Gene similarity networks provide tools for understanding eukaryote origins and evolution. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E1594-603 (2013).

49. R. Jain, M. C. Rivera, J. A. Lake, Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 3801–3806 (1999).

50. P. Puigbò, Y. I. Wolf, E. V. Koonin, The tree and net components of prokaryote evolution. *Genome Biol. Evol.* **2**, 745–756 (2010).

51. R. Liu, H. Ochman, Stepwise formation of the bacterial flagellar system. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 7116–7121 (2007).

52. F. E. Baidouri, C. Venditti, S. Suzuki, A. Meade, S. Humphries, Phenotypic reconstruction of the last universal common ancestor reveals a complex cell, *bioRxiv* (2020). https://doi.org/10.1101/2020.08.20.260398.

53. I. Sela, Y. I. Wolf, E. V. Koonin, Theory of prokaryotic genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 11399–11407 (2016).

54. L. C. S. Antunes, D. Poppleton, A. Klingl, A. Criscuolo, B. Dupuy, C. Brochier-Armanet, C. Beloin, S. Gribaldo, Phylogenomic analysis supports the ancestral presence of LPS-outer membranes in the Firmicutes. *Elife* **5** (2016).

55. D. Megrian, N. Taib, J. Witwinowski, C. Beloin, S. Gribaldo, One or two membranes? Diderm Firmicutes challenge the Gram-positive/Gram-negative divide. *Mol. Microbiol.* **113**, 659–671 (2020).

56. N. Taib, D. Megrian, J. Witwinowski, P. Adam, D. Poppleton, G. Borrel, C. Beloin, S. Gribaldo, Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. *Nat Ecol Evol* **4**, 1661–1672 (2020).

57. E. I. Tocheva, D. R. Ortega, G. J. Jensen, Sporulation, bacterial cell envelopes and the origin of life. *Nat. Rev. Microbiol.* **14**, 535–542 (2016).

58. I. Sánchez-Andrea, I. A. Guedes, B. Hornung, S. Boeren, C. E. Lawson, D. Z. Sousa, A. Bar-Even, N. J. Claassens, A. J. M. Stams, The reductive glycine pathway allows autotrophic growth of Desulfovibrio desulfuricans. *Nat. Commun.* **11**, 5090 (2020).

59. M. C. Weiss, F. L. Sousa, N. Mrnjavac, S. Neukirchen, M. Roettger, S. Nelson-Sathi, W. F. Martin, The physiology and habitat of the last universal common ancestor. *Nat Microbiol* **1**, 16116 (2016).

60. F. L. Sousa, W. F. Martin, Biochemical fossils of the ancient transition from geoenergetics to bioenergetics in prokaryotic one carbon compound metabolism. *Biochim. Biophys. Acta* **1837**, 964–981 (2014).

61. F. L. Sousa, S. Nelson-Sathi, W. F. Martin, One step beyond a ribosome: The ancient anaerobic core. *Biochim. Biophys. Acta* **1857**, 1027–1038 (2016).

62. G. Borrel, P. S. Adam, S. Gribaldo, Methanogenesis and the Wood-Ljungdahl Pathway: An Ancient, Versatile, and Fragile Association. *Genome Biol. Evol.* **8**, 1706–1711 (2016).

63. G. Fuchs, Alternative pathways of carbon dioxide fixation: insights into the early evolution of life? *Annu. Rev. Microbiol.* **65**, 631–658 (2011).

64. T. Nunoura, Y. Chikaraishi, R. Izaki, T. Suwa, T. Sato, T. Harada, K. Mori, Y. Kato, M. Miyazaki, S. Shimamura, K. Yanagawa, A. Shuto, N. Ohkouchi, N. Fujita, Y. Takaki, H. Atomi, K. Takai, A primordial and reversible TCA cycle in a facultatively chemolithoautotrophic thermophile. *Science* **359**, 559–563 (2018).

65. K. Schuchmann, V. Müller, Autotrophy at the thermodynamic limit of life: a model for energy conservation in acetogenic bacteria. *Nat. Rev. Microbiol.* **12**, 809–821 (2014).

66. K. S. Makarova, D. H. Haft, R. Barrangou, S. J. J. Brouns, E. Charpentier, P. Horvath, S. Moineau, F. J. M. Mojica, Y. I. Wolf, A. F. Yakunin, J. van der Oost, E. V. Koonin, Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol.* **9**, 467–477 (2011).

67. E. V. Koonin, K. S. Makarova, Origins and evolution of CRISPR-Cas systems. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20180087 (2019).

68. J. K. Nuñez, P. J. Kranzusch, J. Noeske, A. V. Wright, C. W. Davies, J. A. Doudna, Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat. Struct. Mol. Biol.* **21**, 528–534 (2014).

69. K. S. Makarova, Y. I. Wolf, O. S. Alkhnbashi, F. Costa, S. A. Shah, S. J. Saunders, R. Barrangou, S. J. J. Brouns, E. Charpentier, D. H. Haft, P. Horvath, S. Moineau, F. J. M. Mojica, R. M. Terns, M. P. Terns, M. F. White, A. F. Yakunin, R. A. Garrett, J. van der Oost, R. Backofen, E. V. Koonin, An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **13**, 722–736 (2015).

70. R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. A. Romero, P. Horvath, CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).

71. E. V. Koonin, The origins of cellular life. *Antonie Van Leeuwenhoek* **106**, 27–41 (2014).

72. M. Krupovic, V. V. Dolja, E. V. Koonin, Origin of viruses: primordial replicators recruiting capsids from hosts. *Nat. Rev. Microbiol.* **17**, 449–458 (2019).

73. A. C. J. Roth, G. H. Gonnet, C. Dessimoz, Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* **9**, 518 (2008).

74. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

75. A. Criscuolo, S. Gribaldo, BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).

76. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

77. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

78. A. A. Davín, T. Tricou, E. Tannier, D. M. de Vienne, G. J. Szöllősi, Zombi: a phylogenetic simulator of trees, genomes and sequences that accounts for dead linages. *Bioinformatics* **36**, 1286–1288 (2020).

79. J. Huerta-Cepas, K. Forslund, L. P. Coelho, D. Szklarczyk, L. J. Jensen, C. von Mering, P. Bork, Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).

80. G. A. Coleman, A. A. Davín, T. A. Mahendrarajah, L. L. Szánthó, A. Spang, P. Hugenholtz, G. J. Szöllősi, T. A. Williams, Extended Data for A rooted phylogeny resolves early bacterial evolution, figshare (2021); https://doi.org/10.6084/M9.FIGSHARE.12651074.V9.

81. N. Dombrowski, T. A. Williams, J. Sun, B. J. Woodcroft, J.-H. Lee, B. Q. Minh, C. Rinke, A. Spang, Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. *Nat. Commun.* **11**, 3939 (2020).

82. M. Wilkinson, J. O. McInerney, R. P. Hirt, P. G. Foster, T. M. Embley, Of clades and clans: terms for phylogenetic relationships in unrooted trees. *Trends Ecol. Evol.* **22**, 114–115 (2007).

83. S. M. van Dongen, "Graph clustering by flow simulation," thesis, University of Utrecht, Utrecht, Netherlands (2000).

84. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

85. G. J. Szöllosi, E. Tannier, N. Lartillot, V. Daubin, Lateral gene transfer from the dead. *Syst. Biol.* **62**, 386–397 (2013).

86. H.-C. Wang, B. Q. Minh, E. Susko, A. J. Roger, Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Syst. Biol.* **67**, 216–235 (2018).

87. E. Susko, A. J. Roger, On reduced amino acid alphabets for phylogenetic inference. *Mol. Biol. Evol.* **24**, 2139–2150 (2007).

88. C. Zhang, E. Sayyari, S. Mirarab, "ASTRAL-III: Increased scalability and impacts of contracting low support branches" in *Comparative Genomics* (Springer International Publishing, Cham, 2017)*Lecture notes in computer science*, pp. 53–75.

89. S. R. Eddy, Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).

90. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29-37 (2011).

91. T. Aramaki, R. Blanc-Mathieu, H. Endo, K. Ohkubo, M. Kanehisa, S. Goto, H. Ogata, KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).

92. P. Jones, D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.-Y. Yong, R. Lopez, S. Hunter, InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

93. J. Huerta-Cepas, D. Szklarczyk, D. Heller, A. Hernández-Plaza, S. K. Forslund, H. Cook, D. R. Mende, I. Letunic, T. Rattei, L. J. Jensen, C. von Mering, P. Bork, eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).

94. J. Huerta-Cepas, D. Szklarczyk, K. Forslund, H. Cook, D. Heller, M. C. Walter, T. Rattei, D. R. Mende, S. Sunagawa, M. Kuhn, L. J. Jensen, C. von Mering, P. Bork, eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**, D286-93 (2016).

95. R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, D. A. Natale, The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).

CHAPTER 5

Evolving Perspective on the Origin and Diversification of Cellular Life and the Virosphere

Anja Spang, Tara A. Mahendrarajah*, Pierre Offre*, and Courtney W. Stairs*
*these authors contributed equally to this work

Genome Biology and Evolution, 2022

SUMMARY AND CONTRIBUTIONS

The motivation behind this review was to provide an updated perspective on the tree of life (TOL) within the context of technological advancements and analytical findings of the past couple decades. Our increased understanding of the enormous diversity of the Archaea, Bacteria, and eukaryotes are a result of advancements in genome sequencing technologies that have allowed for sampling across a wider variety of environments and more refined phylogenetic techniques to better place novel lineages. Here we examined how the shape of the TOL, from the deepest split to the branches leading to eukaryotes, has changed in recent years. Debate over whether the current data best supports a three-domain (3D tree) or two-domain (2D tree) topology for the TOL, is centered on the discovery of the Asgard archaea, and their proposed role as the archaeal ancestor of eukaryotes. Advanced sequencing technologies have expanded the categorized diversity of viruses, mobile genetic elements (MGEs), and other parasitic replicators. Recent molecular analyses have shown that viral evolutionary history is intimately linked to cellular evolution, with their replication (genetic) module likely arising in the primordial replicon pool and their morphogenetic (capsid) evolving on multiple occasions, possibly related to interactions with cellular hosts. Furthermore, many viral features of diverse groups are implicated in the evolution of key cellular features. We provide an outlook for how to make further progress in all of these research areas to disentangle and refine hypotheses and open questions about the TOL and viral contributions. In all, greater sampling and sequencing of all major groups in the TOL and the viral realms would provide more information from which to shape our understanding of cellular and viral evolution. The further improvement of evolutionary models would better reflect the processes of gene flow and evolution. Ancestral genome reconstruction provides a window into past metabolisms that can help to link evolution with ecology and Earth history. Advanced cultivation-dependent tools will also be instrumental in understanding the morphology and physiology of important species, including the Asgard archaea, in order to better understand their cell biological features. Furthermore, many ultrasmall cells, which have been shown to be widespread across the biosphere, often exist in symbioses with hosts. Advanced microscopy and imaging can provide a window into understanding the nature of symbiotic lifestyles.

I contributed to writing, revising, and analyzing all elements included in this review. My primary focus included researching, compiling, and summarizing current literature on viruses, including their evolutionary history, diversity, and hypothetical involvement in cellular evolution. I designed the virus section to cover two broad categories: 1) studies analyzing the evolutionary histories of the replication (genetic) and morphogenetic (capsid) modules, and 2) an overview of how morphologic and genetic features of different viral groups reshape our understanding of cellular evolution in prokaryotes and eukaryotes. I helped conceptualize and design the main text figures 1 and 2. I wrote the section on “Viruses and the Tree of Life” and contributed to writing and revision of the entire manuscript at all stages.

ABSTRACT

The tree of life (TOL) is a powerful framework to depict the evolutionary history of cellular organisms through time, from our microbial origins to the diversification of multicellular eukaryotes that shape the visible biosphere today. During the past decades, our perception of the TOL has fundamentally changed, in part, due to profound methodological advances, which allowed a more objective approach to studying organismal and viral diversity and led to the discovery of major new branches in the TOL as well as viral lineages. Phylogenetic and comparative genomics analyses of these data have, among others, revolutionized our understanding of the deep roots and diversity of microbial life, the origin of the eukaryotic cell, eukaryotic diversity, as well as the origin, and diversification of viruses. In this review, we provide an overview of some of the recent discoveries on the evolutionary history of cellular organisms and their viruses and discuss a variety of complementary techniques that we consider crucial for making further progress in our understanding of the TOL and its interconnection with the virosphere.

SIGNIFICANCE

Our review provides a timely overview of how recent methodological progress has allowed an updated view on the tree of life and its connection to the virosphere. It covers topics ranging from last universal common ancestor to last eukaryotic common ancestor and the extant diversity of prokaryotic and eukaryotic life as well as viruses. Furthermore, we summarize current developments in the field that can help to make further progress in our understanding of deep evolution in the coming years.

INTRODUCTION

All cellular life forms (organisms) on Earth can be assigned to one of the major domains—the Archaea, Bacteria, or Eukaryota (hereafter referred to as eukaryotes) (*1*, *2*). Because all organisms have evolved from a shared last universal common ancestor (LUCA) (*3*), the relationship of extant organisms is often depicted within the framework of a tree of life (TOL) (*4–6*). Upon the discovery of the Archaea, it was assumed that the TOL comprises three distinct branches that evolved vertically since LUCA, with the Bacteria on one side of the root and Archaea and eukaryotes forming sister clades on the other side of the root (*2*). However, recent years have witnessed an increasing body of evidence suggesting that eukaryotes, which comprise both uni- and multicellular representatives, have emerged through a symbiosis of an archaeon and a bacterium, that is, through the merging of two branches from within the Archaea and Bacteria, respectively (Fig. 1) (*7–11*). In turn, Archaea and Bacteria are often referred to as primary domains of life while eukaryotes form a secondary domain of life (*12*, *13*). In contrast, viruses are noncellular obligate intracellular parasites that infect all cellular life forms (*14*). Similar to other selfish genetic elements, viruses are generally not considered within the framework of the TOL (*15*), but are an integral part of the biosphere or biological realm (*14*). They also impact genome evolution of cellular life not only through the exchange of genes with their hosts but also through host-parasite coevolution (*16*, *17*). In fact, the prevalence of horizontal gene transfer (HGT) via both mobile genetic elements (MGEs) and viruses but also directly between distinct organisms has to some extent questioned the concept of a TOL, which may be more correctly represented as a network including both vertical and horizontal branches (*4*, *5*, *18*). Yet, despite this component of horizontal genome evolution, the "statistical" TOL has remained a useful concept for understanding life's diversification (*6*, *19*).

Recently, the application of cultivation-independent metagenomic and single-cell genomic techniques has improved our knowledge of microbial and viral diversity and, in turn, our view of the TOL (*20*) and its connection to the virosphere (*21*). For example, during the past decade a plethora of previously unknown archaeal and bacterial taxa (e.g., reviewed in (*22–24*)) have been described, including various lineages of high-taxonomic rank at the phylum and class-level (*20*, *25*, *26*). Furthermore, progress has been made with regard to our understanding of the origin of eukaryotes (*11*) as well as their subsequent diversification (*27*). Genomics approaches have also transformed our knowledge on the vast diversity of viruses (*28–34*), their putative host taxa (*35–38*), and origins (*39*).

In the following, we will provide an updated perspective of the TOL and virosphere by focusing on selected key findings. Furthermore, we describe a variety of research approaches, which we consider important for making further progress on our understanding of the history of life on Earth.

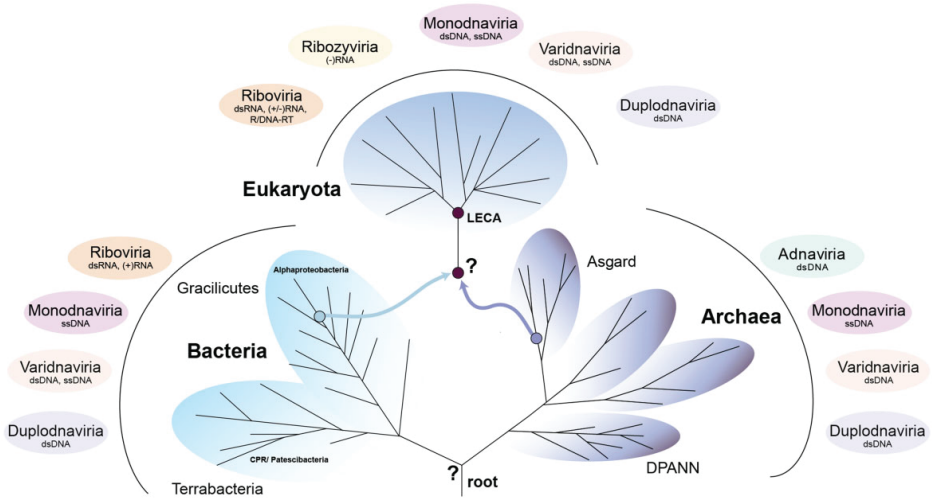


Fig. 1 | Tree of cellular life (TOL) and connection to the six major realms of viruses. The tree is a schematic representation of the relationship of the major domains of life, comprised of the primary domains of Archaea and Bacteria and the secondary domain of Eukaryota. The assumption that Archaea and Bacteria form separate domains of life is dependent on the placement of the root between those domains, though this hypothesis remains to be validated. Although the node separating the DPANN (acronym referenced in text) from all other archaeal clades has been suggested to be the most ancestral split on the archaeal branch, the CPR (acronym referenced in text) most likely represents a more recently evolved sister-clade of the Chloroflexota (*64*). Current data support an origin of the eukaryotic cell through a symbiosis between an ancestral member of the Asgard archaea (also Asgardarchaeota) (purple arrow) and Alphaproteobacteria (blue arrow), though the timing of the mitochondrial acquisition is debated and the events leading to LECA are poorly resolved. On the outside of the TOL, we illustrate the connection of the three cellular domains with virus representatives belonging to either of the six major viral realms, the Riboviria, Monodnaviria, Varidnaviria, Duplodnaviria, Adnaviria, and Ribozyviria (*21*, *183*). The latter two realms are restricted to the Archaea or eukaryotes, respectively. The Riboviria have so far only been found associated with Bacteria and eukaryotes, whereas all other realms include members infecting cellular organisms across the TOL. LECA, last eukaryotic common ancestor.

THE PRIMARY DOMAINS OF LIFE AND DEEP ROOTS OF THE TOL

The nature of LUCA and the emergence of the two primary domains of life are some of the most fundamental unknowns in our understanding of life's evolution. Archaeal and bacterial cells are distinguished by major differences in their cell lipid membrane and use of contrasting molecular machinery, including for the replication, and processing of genetic information. Although a wide variety of hypotheses have been proposed to explain the distinct cell membranes of bacteria and archaea and the early evolution of their metabolism, these remain controversial and progress has been constrained by the limited availability of relevant data (*40–43*). It is generally assumed that the root in the TOL separates Archaea and Bacteria as inferred based on the use of ancient paralogous gene families for rooting (*44–47*) and genome networks (*48*) (Fig. 1). Yet, the accurate placement of the root is challenging and prone to phylogenetic artifacts and alternative roots, such as within Bacteria (*49*, *50*), have not been

formally ruled out (*51*). Further, it has recently been suggested that the branch separating the primary domains of life may be shorter than in previous estimates (*52*). However, it was subsequently shown that the reduced estimate of the Archaea/Bacteria branch length most likely results from inter-domain gene transfers and, in agreement with earlier work (*19*, *20*), that the longest branch in the TOL lies between Archaea and Bacteria (*53*, *54*) (note that these analyses did not consider extremely fast-evolving symbionts and parasites). Improved phylogenetic models, the integration of genomic data from the diversity of recently discovered taxa as well as the use of novel approaches for rooting, such as gene tree-species tree reconciliations, for example, (*55–57*) (see below), will help to determine whether this branch indeed represents the deepest split in the TOL.

Particularly, the discovery of two previously unknown and potentially deep-branching microbial radiations in the Bacteria and Archaea, the so-called DPANN archaea (*58*, *59*) and the bacterial Candidate Phyla Radiation (CPR or Patescibacteria) (*60*), respectively, has provided important data for readdressing the deep roots of microbial life and the placement of the archaeal and bacterial roots (*61–65*). The DPANN group (acronym referring to its first described member lineages, the Diapherotrites, Parv-, Aenigm-, Nano-, and Nanohaloarchaeota) now includes more than eight distinct archaeal phyla (*26*) that group together with Nanoarchaeota, an archaeal clade represented by the ultrasmall and ectosymbiotic archaeon *Nanoarchaeum equitans* (*66*). Representatives of DPANN have small genomes and cell sizes, are characterized by restricted anabolic and catabolic capabilities, and include obligate ectosymbionts some of which have been cultivated in coculture with their hosts belonging to the Halobacteriota, Thermoproteota, and Thermoplasmatota (*66–75*). Indeed, symbiotic lifestyles have been suggested to represent a common feature of genome-reduced members of the DPANN (*62*). Likewise, members of the CPR, which also include various lineages of high taxonomic rank, share several genomic features with the DPANN archaea, such as small cell and genome sizes, a limited metabolic potential and potential dependency on partner organisms (*62*). In line with this, two representatives of this group, that is, members of the Saccharibacteria and Absconditabacteria, have been successfully enriched as symbionts in coculture with their respective actinobacterial and gammaproteobacterial hosts (*76–78*). It seems that the level of host specificity differs significantly between different representatives of the DPANN and CPR. For instance, although the most genome-reduced members of the DPANN, such as *N. equitans*, seem unable to switch between different host strains (*79*), members of the Micrarchaeota infect hosts belonging to different archaeal phyla and comprise strains that can grow in coculture with hosts belonging to different genera (*70*, *71*, *75*). Furthermore, it seems that at least DPANN may also include free-living members such as the Altiarchaeota (*80*) or members, which, in spite of certain auxotrophies, do not require permanent physical contact with potentially interacting partners (*81*, *82*).

Initial phylogenetic analyses have recovered both the CPR (*60*) and DPANN (*58*, *59*) as monophyletic and early diverging branches in the TOL (Fig. 1), but these findings are being debated (*83*, *84*). In particular, several authors have raised the concern, that the deep and

monophyletic placement of DPANN and CPR lineages may be the result of phylogenetic artifacts (*85–89*) such as long-branch attraction, that leads to the erroneous grouping of fast-evolving taxa in a monophyletic clade as well as their attraction to a distant outgroup (*90*, *91*). For example, previous studies have revealed that genomes of other symbionts (e.g., obligate intracellular bacterial endosymbionts) indeed experience faster evolutionary rates, have compositional biases and form long branches in phylogenetic trees (*92*, *93*). In turn, elucidating the phylogenetic placement of the symbiotic CPR and DPANN has proven challenging and requires careful phylogenetic approaches implementing, among others, careful marker gene and taxon selection approaches and/or the use of complex models of evolution that account for differences in evolutionary rates across sites and lineages (*53*, *64*, *94*). Furthermore, such analyses benefit from taking into account potentially increased rates of HGT between symbionts and their hosts (*94*).

Recently, outgroup-free rooting methods have been applied to assess the placement of CPR and DPANN in the TOL. For instance, (*64*) have used a gene tree—species tree reconciliation approach (*55–57*) to root the bacterial tree and reconstruct the proteome of the last bacterial common ancestor. Interestingly, and in contrast to several earlier studies, this has revealed that the CPR most likely represents a more recently evolved monophyletic sister-lineage of the Chloroflexota (*64*) rather than an early diverged bacterial clade (*60*) (Fig. 1). Thus, CPR members seem to be derived from more complex ancestors with their small genomes being a result of genome-streamlining processes (*64*). In agreement with this, a recent analysis aiming to resolve the evolution of cell envelopes in Bacteria not only indicated the ancestry of didermy with several independent transitions to monoderm phenotypes but also supported a sisterhood relationship of Chloroflexota and CPR nested within Terrabacteria (*63*). Finally, the careful assessment of marker genes for multidomain phylogenies has further confirmed this derived placement of the CPR (*53*).

In contrast, several recent studies have provided support for the “clanhood” of DPANN in unrooted phylogenies, their characteristic set of genes and their placement as an early radiation on the archaeal branch of the TOL raising the possibility that DPANN clades may have evolved in parallel with their host lineages over much of evolutionary time, see for example, (*53*, *61*, *94–96*), (Fig. 1). However, conflicting results regarding the placement of certain putative DPANN clades remain (*89*). Furthermore, it is important to note that the exact placement of the root in the archaeal tree is not yet fully resolved and could be located between two distinct DPANN clades, thus leaving open the possibility that DPANN are paraphyletic (*94*, *96*). Further analyses, such as the application of gene tree-species tree reconciliations applied to a larger set of representative archaeal genomes will help to test current hypotheses on the early divergence of DPANN. Finally, a reliable interpretation of the early evolution of cellular life, the features of the last universal common ancestor, and the relationship of DPANN and CPR, hinges on the accurate placement of the universal root (*51*).

ORIGIN OF THE EUKARYOTIC CELL FROM PROKARYOTIC ANCESTORS

The origin of the eukaryotic cell represents one of the most significant and at the same time debated events in life's evolution. Over the years, a variety of eukaryogenesis models have been put forth, which can be broadly categorized into symbiogenetic and autogenous models, discussed in several comprehensive reviews (*7*, *10*, *97*, *98*). Although autogenous models assume the vertical evolution of a protoeukaryotic lineage from a root shared with the archaeal and bacterial line of descent, symbiogenetic models suggest that the origin of the eukaryotic cell is a result of a merger of members of at least two distinct microbial lineages belonging to the Archaea and Alphaproteobacteria (*9*, *99*) (Fig. 1).

Recently, the genomics-based discovery of the Asgard archaea (*100*, *101*) (also referred to as the phylum Asgardarchaeota (*26*)), has provided important data shedding new light on the origin of the eukaryotic cell. Asgard archaea were originally described to comprise the Loki-, Thor-, Odin-, and Heimdallarchaea (*100–102*), but are now known to include a variety of additional clades (*103–108*). Notably, phylogenetic analyses have revealed that the Asgard archaea comprise the closest archaeal sister lineage of eukaryotes (*101*, *105*, *107*) and thereby provided increasing evidence for the evolution of eukaryotes from within the Archaea (*12*, *13*) (Fig. 1). But although there is strong support for the monophyly of Asgard archaea and eukaryotes, the exact placement of the eukaryotic branch relative to the various Asgard lineages varies depending on data set composition and evolutionary models used (*13*, *101*, *105*). Expanded sampling of Asgard diversity combined with careful phylogenetic analyses, is likely to provide improved resolution of branching orders and will allow to pinpoint the closest sister-lineage of eukaryotes more precisely.

In agreement with phylogenetic evidence, comparative analyses of the Asgard archaeal genomes have revealed the presence of so-called eukaryotic signature proteins (ESPs) (reviewed in (*11*, *23*, *109*)), that is, proteins that were previously thought to be absent from prokaryotic genomes. Notably, these ESPs are homologous to proteins integral to the functioning of complex eukaryotic cells and comprise essential building blocks of the ESCRT (endosomal sorting complex required for transport) system, ubiquitin, trafficking, and informational processing machineries as well as the cytoskeleton (*100*, *101*, *105*). Although the function of these proteins in Asgard archaea remains to be elucidated, the heterologous expression and structural analyses of some of these proteins such as profilins and gelsolins have revealed that they are functionally equivalent to their eukaryotic homologs and suggests that a regulated actin cytoskeleton precedes eukaryogenesis (*110–112*).

Because even high quality metagenome assembled genomes (MAGs) (i.e., completeness >90% and contamination <5%, according to (*113*)) usually do not assemble into complete genomes and may contain a low amount of contamination from genomes of other community members or closely related strains, some studies have questioned the reliability of the Asgard archaeal MAGs and in particular raised concerns as to whether ESPs may represent contamination rather than being genuine genomic signatures (*114–116*). However, various lines of evidence during

the past years have supported the existence of Asgard archaea, the emergence of the archaeal ancestor of eukaryotes from within this group as well as the presence of ESPs as part of their coding potential: among others, ESPs are encoded within a prokaryotic genomic context, lack introns characteristic of many eukaryotic genes, and are significantly divergent from eukaryotic homologs to exclude contamination (*100*, *101*, *117*). Furthermore, Asgard MAGs have now been reconstructed from a large variety of metagenomes from different environmental samples all over the world and by many different research groups, yet show consistent genomic signatures across the various member clades (*104–106*, *108*, *118–120*). Even though the presence/absence pattern of ESPs across Asgard archaea is variable and indicates a complex history of ESP evolution involving duplications, differential loss, and transfers, the shared set of ESPs within specific taxon-level (e.g., class-level) lineages is very consistent and provides strong evidence for ESPs representing genuine signatures of Asgard proteomes (*105*). In line with this, the successful enrichment of the first representative of the Asgard archaea, *Candidatus* Prometheoarchaeum syntrophicum has not only proven the viability of members of this group but also allowed the reconstruction of the first complete genome of a Lokiarchaeote with a characteristic and consistent set of ESPs (*121*). Finally, initial microscopy analyses have provided insights into the cellular features of extant members of the Asgard archaea including cellular protrusions (*121*, *122*) and revealed the spatial separation of genomic DNA and ribosomes in certain representatives (*122*).

The analysis of the genomic repertoire of the Asgard archaea has not only enabled predictions of their extant metabolic characteristics but also provided a first baseline to refine symbiogenetic eukaryogenesis models, which predict a syntrophic interaction as an important initial driver for cell–cell interactions (*9*, *105*, *121*, *123*), and represent an extension of the Hydrogen (*124*) and Syntrophy (*125*) Hypotheses. However, more detailed models hinge on resolving the exact placement of the eukaryotic and mitochondrial branches relative to the Asgard archaea (*101*, *105*) and Alphaproteobacteria (*99*, *126–128*), respectively, as well as the cellular and metabolic features of these ancestors. Additionally, controversies remain with regard to the timing of the events during eukaryogenesis, that is, the timing of the mitochondrial acquisition, the evolution of an endomembrane system as well as the establishment of a nucleus, for example, (*11*, *129–134*) (Fig. 1). Finally, the extent to which additional microbial lineages and/or viruses (see below) have contributed to the eukaryotic proteome are still to be determined. Phylogenomics analyses have for example provided support for the hypothesis that the genomic repertoire of eukaryotes was shaped through genetic input from Bacteria other than Alphproteobacteria (*135–139*) as well as by viruses, for example, (*140–143*). Furthermore, a recently proposed updated symbiogenetic model on the origin of the eukaryotic cell has implicated the potential involvement of an additional bacterial lineage (i.e., a Deltaproteobacterium) during eukaryogenesis (*9*).

The combination of novel techniques in phylogenetics with cell biological and cultivation approaches (see below) will help to address those conflicting hypotheses of the origin of the complex eukaryotic cell from its prokaryotic ancestors and continue to illuminate the timing of the events during eukaryogenesis (*11*, *144*).

EUKARYOTIC DIVERSITY AND THE LAST EUKARYOTIC COMMON ANCESTOR

Even though various aspects of eukaryogenesis remain enigmatic, our knowledge of the last eukaryotic common ancestor (LECA) (reviewed in (*11*)) and its subsequent diversification has grown substantially in recent years, enabled by a tremendous increase in our sampling of extant eukaryotic diversity. Indeed, although the majority of formally described eukaryotes are multicellular and fall into two phylogenetic groups: Archaeplastida (plants and algae) and Opisthokonta (animals and fungi), it is now clear that the bulk of phylogenetic diversity of eukaryotes is composed of unicellular representatives including "protists" and algae (Fig. 2). Major advances in cultivation-dependent (*27*) and cultivation-independent (*145*) methods including symbiosis-aware strategies (*146*) for generating sequence data combined with sophisticated bioinformatic tools for genome assembly, gene annotation, and phylogenomic inference have been critical for the genomics-driven exploration of eukaryotic biodiversity. In particular, the last decade has witnessed the discovery of numerous kingdom- and phylum-level lineages and confidently placed those in the eukaryotic TOL (Fig. 2), for example, Rhodelphia (*147*), Picozoa (*148*), Anaeramoebae (*149*), and "CruMs" (*150*) (Collodictyonids, Rigifilids, Mantamonads). Sequence data has also been collected from lineages that have no clear phylogenetic position including *Ancoracysta twista* (*151*), Hemimastigophora (*152*), Ancyromonadida (*153*), and Malawimonadida (*154*) that might each represent phylum- (or higher-) level taxonomic ranks.

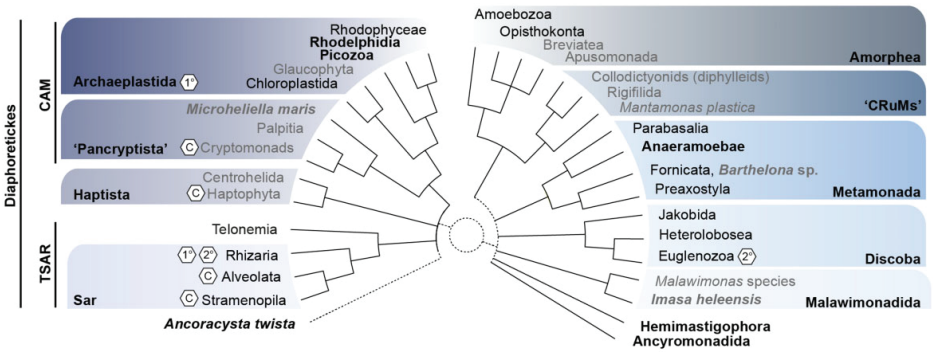


Fig. 2 | Schematic representation of the phylogenetic diversity of eukaryotes. Groups with taxonomic rankings of phylum level or higher are shown in black (according to (*321*) and references in text). Select lineages or organisms that have been recently discovered and placed in the eukaryotic TOL are shown in bold. Eukaryotic supergroups are colored for clarity. Lineages with one or more representative with a primary (1°) secondary (2°) or complex red (C) plastids are indicated with hexagons based on (*182*). Sar, Stramenopila-Alveolata-Rhizaria; TSAR, Telonemia+SAR (*322*), CAM (*323*), Cryptista-Archaeplastida-*Microheiliella maris*; "CRuMs," Collodictyonids, Rigifilida, *Mantamonas plastica*.

Supported by these new data, numerous lines of evidence suggest that LECA dated to the Proterozoic (ca. 1.9–1.6 billion years ago) (*155–157*) and was characterized by a nucleus and nuclear pores, linear chromosomes with telomeres, genes with spliceosomal introns, complex

RNA processing, and regulatory mechanisms, an elaborate endomembrane system (including a Golgi apparatus, endosomes, lysosomes, and peroxisomes), mitochondria, bacterial-type lipids as well as a complex cell cycle (extensively reviewed in (*158*) and (*11*)). Some analyses predict that the LECA proteome was already quite complex with many orthologs (~10,000) tracing their origin to LECA (*159*), though many details regarding components of the various cellular and molecular machineries remain to be further illuminated. One current limitation lies in the unresolved placement of the root in the eukaryotic tree. Depending on gene set and methodology used, the root of the eukaryotic tree has been inferred between Discoba and other eukaryotes (*160*), between Diaphoretickes + Discoba and Amorphea + CruMs + Malawimonads (*161*) or between Opisthokonta and all other eukaryotes (*162*, *163*). Therefore, the best-studied eukaryotes on which various previous LECA inferences are based, represent derived clades on either side of the putative root: the Archaeplastida within Diaphoretikes and Opisthokonta within Amorphea. It is conceivable that genes conserved in either of these lineages may not necessarily trace their origins back to LECA. For example, a recent review by (*164*) put forth a new term defining hidden ancient homologs as "jotnarlogs" that are shared across eukaryotic biodiversity exclusive of the "model system" lineages. They show that these jotnarlogs are highly relevant for our understanding of the earliest steps in eukaryotic evolution and, among others, comprise proteins mediating fundamentally eukaryotic processes including mitochondrial division (*165*) and membrane trafficking (*164*). In turn, prospective analyses that make use of the increased sampling of eukaryotic genomic diversity will be crucial to further improve our knowledge on the nature of LECA as well as the root placement in the eukaryotic TOL.

Although most modern eukaryotes share key cellular features, the recent discovery of novel eukaryotic representatives forming distinct branches in the eukaryotic tree have revealed interesting insights into eukaryotic metabolic and cellular diversity. For example, although the alphaproteobacteria-derived mitochondria in extant aerobic eukaryotes house the respiratory chain that couples ATP biosynthesis to the reduction of oxygen, in some anaerobic animals and fungi, the respiratory chain uses alternative electron acceptors to oxygen in order to synthesize ATP, often by "tinkering" with existing cellular systems to synthesize anaerobiosis-specific cofactors or by encoding anaerobiosis-specific proteins (*166*, *167*). Further, many anaerobic protists have lost most, if not all, respiratory capabilities and instead couple ATP biosynthesis to fermentative H_2 production within so-called mitochondria-related organelles (MROs) (*166–168*). Some representatives, such as *Monocercomonoides,* have lost their MROs (*169*), and/or mitochondrial genomes (*168*) entirely. The genetic origins of the anaerobic metabolism of MROs remains a widely debated topic (see, e.g., (*133*, *138*, *170–173*)).

Photosynthesis is a widespread trait across the tree of eukaryotes with representatives in Stramenopila, Alveolata, Rhizaria, Haptista, Pancryptista, Archaeplastida, and Discoba. Primary plastids, derived from the engulfment of an ancestral photosynthetic cyanobacterium with the closest present day relative likely being *Gloeomargarita lithophora* (*174*, *175*), have evolved at least once on the tree of eukaryotes in the Archaeplastida (*173*) between 2.1 and

1.6 (*176*, *177*) or 1.8 and 1.1 billion years ago (*157*). There is at least one additional candidate of a primary photosynthetic organelle in eukaryotes in the Rhizarian *Paulinella chromatophora* (*178*, *179*). This amoeba houses a specialized organelle called the chromatophore that has its own genome and is thought to have evolved from an ancestral endosymbiont of the *Synechococcus/Prochlorococcus* clade (*180*) roughly 90–140 Ma (*181*). The chromatophore provides a rare opportunity to study the early stages of endosymbiosis having occurring nearly 1 billion years more recently than the primary plastids of Archaeplastida. Other eukaryotes, that is, heterotrophic protists, have acquired secondary or higher order plastids through serial endosymbiosis events, reviewed in (*182*). These higher-order plastids are often surrounded by three or four membranes and, in at least three separate lineages, retain the nuclei (dubbed the nucleomorph) from the engulfed endosymbiotic algae (*182*). In these cells, there can be as many as four distinct genomes derived from the host nucleus, host mitochondrion, plastid, and nucleomorph. Continued investigations comparing the origin of the gene content and cell biology of these diverse and complex algal lineages as well as phylogenetic and molecular dating approaches will help in identifying the mechanisms necessary for enabling endosymbiosis events and help to further improve our understanding of their timing throughout eukaryotic diversification (*177*).

VIRUSES AND THE TREE OF LIFE

MGEs are semiautonomous replicative genomic entities that are ubiquitous in the natural environment and believed to be an intrinsic part of cellular evolution (*183*). They include viruses which may encode one or more proteins comprising the viral particle (virion) encasing the genome of the respective MGE (*183*). Categorically, viruses are believed to be the most abundant biological entities on the planet, shaping ecological and evolutionary components of the biosphere (*39*). The diverse characteristics of MGEs stratify the semiautonomous replicative genomic entities or replicator groups, blurring the boundaries between the major categories within the replicator space, with the Virosphere defined at its core by the *Orthovirosphere*, followed by the *Perivirosphere*, and the remaining replicators falling within the periphery (*183*).

Recent evolutionary insight has classified the core of the virosphere, that is, the *Orthovirosphere*, into six major realms, the *Riboviria*, *Varidnaviria*, *Duplodnaviria*, *Monodnaviria*, *Adnaviria*, and *Ribozyviria* (*183*), comprising many but not all viral families (Figs. 1 and 3). Apart from the *Ribozyviria*, which has been identified in specific vertebrates, all realms are believed to have emerged before or near the origination of the last universal cellular ancestor (LUCA) (*21*, *183*). To fully understand the roles viruses played during the earliest stages of the evolution of cellular life, studies have sought to understand the origins of key viral components. Generally, viral genomes are unified by two core modules: a module that encodes the proteins responsible for genome replication (the replication module) and a module that encodes the proteins that form the virion particle that encapsulates the genome (the morphogenetic module) (*39*). Despite great viral diversity, most replication modules can be

captured by four hallmark replication protein families: the RNA-dependent RNA polymerase, the reverse transcriptase, the protein-primed family B DNA polymerase, and the rolling-circle endonuclease (*39*). All of these share the common ancient RNA-recognition fold and importantly, have minimal to no close sequence identity with replication proteins from cellular organisms. Conversely, investigation into the origins of the capsid proteins that comprise the virion suggests descent from protein families from cellular ancestors, specifically those involved in carbohydrate- or nucleic acid binding (*39*). These findings are the foundation of the proposed chimeric model of viral evolution which describes the emergence of the replication module from the primordial replicon pool, with the morphogenetic module evolving on several different occasions through life's history by acquisitions of structural proteins from hosts (*39*). Notably, recent structural and genomics studies into the diversity of archaeal viruses have revealed an abundance of archaea-specific viruses that share no genetic or structural similarity to bacterial and eukaryotic counterparts (*184*, *185*) and cannot currently be assigned to any of the viral realms (Fig. 3). Beyond unique morphologies across the archaeal viruses, the archaea-specific *Adnaviria* possess a morphogenetic module composed of a capsid protein with a distinct fold not captured by viruses in the other two domains (*183*). These findings underscore the need for further exploration into the diversity, structure, and function of archaeal viruses.

Fig. 3 | The diversity of the core virosphere and its links to bacterial, archaeal, and eukaryotic hosts. For each viral realm, we depict the diversity of viral families that have representatives infecting members either the Bacteria, Archaea, or Eukaryota, respectively. Asterisk: for eukaryotic viruses assigned to the Riboviria, we report orders instead of families. The shapes represent a small selection of characteristic morphologies seen within certain viral realms. The information on viral families comprising the various realms is derived from the ICTV database (https://talk.ictvonline.org/files/master-species-lists/), that is, ICTV Master Species List 2020.v1.xlsx. (*21*, *183*).

Viruses and other MGEs are generally not considered part of the TOL (*186*), however the nature of their replication and propagation mechanisms have linked them to critical components of cellular genome dynamics and evolution. Recent efforts have tried to connect the deep origins and diversification of viruses to the earliest transitions in the TOL and diversification of cellular life (*47*, *187*, *188*). Parasitic replicators play important roles in host-parasite coevolutionary dynamics and the evolution of host genomes (*189*) and have been placed at the centre of debates regarding eukaryotic evolution and diversification (*188*, *190–194*). Particularly the discovery of eukaryotic NucleoCytoplasmic Large DNA viruses (NCLDVs), also referred to as giant viruses (*195*), has sparked debates on the boundaries between viruses and cellular organisms as well as raised questions regarding their origins, relationship to cellular life and role in the origin of the eukaryotic cell. NCLDVs comprise members with unique features among viruses including genome sizes that resemble those of some free-living microorganisms, the presence of genes for DNA maintenance including repair, replication, transcription, and

translation, complex metabolic capabilities, cytoskeleton components, as well as other signature proteins of complex eukaryotic cells, all of which were originally thought to be confined to cellular life (*32*, *196–203*). Some representatives replicate within viral factories, that is, intracellular compartments in which viral components are localized and that may be enclosed by membranes (*204*, *205*), and can be parasitized by their own virophages (*206*). But although those characteristics have originally been suggested to indicate that NCLDVs may form a separate branch within the TOL (*195*), careful phylogenetic analyses have subsequently shown that NCLDVs have acquired hallmark cellular genes through HGT by their hosts and evolved gigantism multiple times (*207–210*), validating the distinction of viruses and cellular life (*14*, *15*, *186*, *211*). Viruses and in particular NCLDVs have also been hypothesized to have played a role in the origin of the nucleus due to the ability of some representatives to assemble viral factories reminiscent of eukaryotic nuclei (*212*). However, the direct involvement of a virus in the origin of eukaryotic organellar complexity remains debated (*213*) and viral factories, including those established by certain *Pseudomonas* phages enclosed by a proteinaceous shell (*214*), likely represent analogous structures to eukaryotic nuclei. Nevertheless, viruses and/or MGEs have been found to have shaped the eukaryotic proteome early on including through virus-to-host HGT (*188*, *192*). For example, the mitochondrial single-subunit RNA polymerase (ssRNAP) has been suggested to be derived from T-odd phages (*140–142*) and eukaryotic telomerases, that ensure the replication of linear chromosomes, are likely derived from a Penelope-like retroelement reverse transcriptase (*190*). The finding of widespread endogenization of viral genomes, including those of NCLDVs, into eukaryotic host genomes highlights a potentially important strategy underlying virus-to-host HGTs (*193*, *215*). Thus, to further disentangle the sources of the eukaryotic proteome and cellular features, prospective phylogenetic analyses benefit from taking into account the wide diversity of viral in addition to prokaryotic genome data (*188*). In this regard, it is particularly noteworthy that recent metagenomics approaches (some only available as preprints so far) have identified a suite of viruses likely infecting Asgard archaea and belonging to different viral realms (*107*, *216–218*). The genomic and experimental analysis of these and other novel viruses may help to test hypotheses on the features and impact of MGEs in the earliest transitions and diversification of eukaryotic cells.

Taken together, a better understanding of the TOL and major evolutionary transitions hinges on the continued exploration of the virosphere combined with improved phylogenomics and network analyses that allow illuminating the impact of viruses and other MGEs on cellular evolution.

HOW TO MAKE FURTHER PROGRESS

Making further progress in our understanding of the TOL and resolving the phylogenetic placement of taxa near key evolutionary branching points requires advances within a wide range of research topics, which we summarize below ((*219*), Fig. 4).

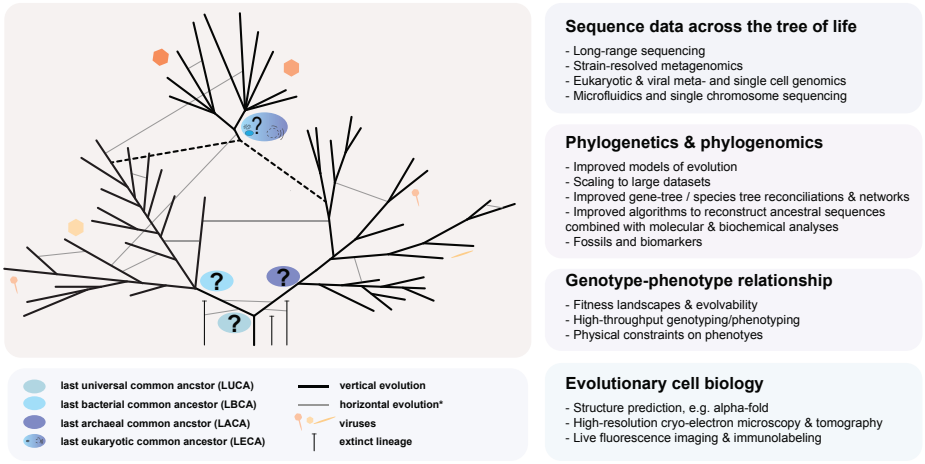


Fig. 4 | Schematic representation of TOL highlighting key questions and approaches to further illuminate cellular evolution and its connection to viral evolution. See text for more details. Asterisks: please note that horizontal evolution has been estimated to be much more prevalent than indicated in the schematic tree.

SEQUENCE DATA ACROSS THE TOL

The availability of molecular sequence data for appropriate and extensive taxa sets is a key factor for the reconstruction of congruent phylogenies and understanding life's evolutionary history in general (*220*). Advances in sequencing and data processing techniques have considerably expanded the set of genomes from uncultivated organisms across the TOL and led to a large set of single-cell and metagenome-assembled genomes (SAGs, MAGs) (*30*, *221–224*). However, the quality of these SAGs and MAGs differs widely (*113*) and, thus far, rarely provide resolution on single strain level. Current developments of hybrid metagenome assembly methodologies combining both short and long DNA sequence reads (*225*, *226*), innovative genome scaffolding approaches using chromosome conformation capture techniques (*200*), and sophisticated (meta)genome assembly computer software (e.g., (*226–228*) for review) are promising avenues to obtain high quality strain-resolved MAGs (*229–231*) including their CRISPR loci as well as ribosomal RNA operon(s). Such improved metagenomics-driven analyses are also valuable not only for expanding the known diversity of DNA viruses (*28–30*, *32–34*), but also to link putative viral genomes to their potential hosts through matching CRISPR spacers (*232*); an approach recently used for the identification of viruses infecting Asgard archaea (*216–218*). Considering the complexity of viral populations, a perhaps even more promising approach relies on improved long-range sequencing technologies and was recently used to obtain complete viral genomes without the need for assembly and binning (*31*).

In contrast to prokaryotes and viruses, many lineages of eukaryotes, and especially microbial representatives, remain only sparsely sampled, which considerably limits our understanding of the early evolution and diversification of these organisms (*233*). Only a small number of protists have been enriched in culture and metagenomic approaches targeting uncultivated protists directly are difficult to implement due to the unique and complex genomic features of

many representatives (*234*), which poses challenges for genome assembly and metagenomic procedures. Further, it should be emphasized that establishing methods for cultivation (or single-cell isolation), nucleic acid isolation, and sequencing from understudied eukaryotes in and of itself is not trivial and requires years of optimization before data analysis can begin (*27*). Many protists harbor symbionts and/or can only be cultivated with other microbes thereby making most protist sequencing projects mini-metagenomics initiatives. Assuming high-quality genomic or transcriptomic data sets can be obtained, the next major obstacle is gene prediction. For genome projects, the nonuniform sequence composition across the genome and the complex architecture of eukaryotic genomes (i.e., large intergenic regions, introns) is a challenge for metagenomic "binning" and gene prediction tools, respectively. Although recent advances in assembling eukaryotic genomes and predicting gene content from complex samples (e.g., nonaxenic cultures or environmental samples) will help in overcoming these obstacles, e.g., (*235*) and (*200*)). Finally, the lack of high-quality reference annotations from diverse eukaryotic representatives, large number of paralogues, and high proportions of lineage or organism-specific putative protein-coding genes in eukaryotic genomes (up to 60% (*236*)) can impede clustering of orthologous groups and poses challenges for the accurate inference of gene history evolution.

PHYLOGENETICS AND PHYLOGENOMICS

Ways to resolve incongruences and uncertainties in phylogenies inferred with state-of-the-art phylogenetic and phylogenomic approaches have been reviewed recently (*220*, *237*) and will not be extensively discussed. These strategies include, among various others, the development of models of DNA and protein sequence evolution that better capture the processes by which molecular sequences evolve and adequately deal with sources of systematic error (i.e., nonphylogenetic signal) in sequence data: for example, see the recent development of heterotachy mixture models (*238*). Much of our understanding of the evolutionary history of life mainly derives from analyses of multigene concatenations based on a limited set of universally conserved single-copy marker genes (see, e.g., (*53*, *54*). Elucidating ancient divergences is challenging and requires the use of metrices to assess confidence in tree topologies and bipartitions. However, classical metrices such as the bootstrap, originally designed for single gene trees, have the tendency to overestimate confidence in bipartitions when the analyses are based on long alignments from multigene concatenations (*239*). In turn, it is valuable to explore improved measures to assess confidence in tree and branching patterns (*240*), such as, for example, the recently developed internode and tree certainty metrices (*53*, *241*). Furthermore, although key to inferring phylogenetic relationships of taxa, multigene concatenations are insufficient to reconstruct the evolution of genomes, which not only results from substitutions but also from gene and genome rearrangements, duplications and the loss and gain of new genes (*242*, *243*). Novel methodologies, capable of capturing simultaneously the vertical and horizontal components of genome evolution such as phylogenetic networks (*244*), topological data analyses (*245*, *246*), as well as gene tree-species tree reconciliation methods (*55–57*, *247*), open up new perspectives toward integrating data from viruses, and other genetic elements as well as providing a deeper understanding of gene family evolution

including both vertical and horizontal components, across the TOL. For instance, reconciliation methods rely on a model to describe gene tree evolution involving originations, duplications, transfers, and losses under a given species tree and allow to determine the probability of any protein family at any given node in a tree (*61*, *64*). Furthermore, such approaches can be used to determine the likelihood of certain root positions in the absence of a remote outgroup (*61*, *64*), which, if available, can cause phylogenetic artifacts such as long branch attraction (*90*, *91*). The modeling of reticulate evolution has recently also been shown to allow dating the TOL (*248*, *249*), which previously solely relied on the scarce fossil and biomarker record available for the early steps of microbial evolution. Together, this can greatly enhance the understanding and timing of the evolutionary trajectories of life.

RECONSTRUCTION OF ANCESTRAL SEQUENCES AND GENOMES

Progress in the sequencing and assembly of ancient DNA has been successfully applied to reconstruct the genome sequence of organisms (*250–253*) including microorganisms (*254–256*) that existed up to hundreds of thousands years ago (i.e., allochronic reconstruction). However, such data is scarce; thus genes, proteins, and genomes of ancestral organisms are predominantly inferred from the sequence of extant taxa using so-called ancestral state reconstruction methodologies (i.e., synchronic reconstruction) (*257*). This includes both ancestral (gene) sequence (*258–261*) and genome reconstruction approaches such as gene tree-species tree reconciliations (see above) (*55–57*, *61*, *64*, *247*). In turn, features of ancestral organisms and the direction of evolutionary change can be investigated simultaneously.

Progressing further in our knowledge of the features of ancestral organisms involves "resurrecting" those life forms or, at least, some of their proteins (*262–264*) before characterizing them using molecular, biochemical, and biophysical approaches. Although this has been successfully undertaken for several types of proteins and protein complexes (*265–268*), features of ancestral proteins and protein complexes thought to have played roles in major evolutionary transitions remain largely unknown. In contrast, the "de novo synthesis" of minimal, ancestral cells, still poses significant challenges (*269*).

EVOLUTIONARY CELL BIOLOGY

Reconstructing and understanding the evolution of the ultrastructural complexity of cells and their components throughout the TOL and, most notably, during eukaryogenesis, requires linking gene and genome sequences to protein structures and cellular features. Although the intracellular organization of bacterial and archaeal cells has long been thought to be relatively simple, tremendous advances of microscopy techniques and image analyses now allow probing the cells of these organisms with sufficient resolution to reveal their cytological features in unprecedented detail (*270*). Cryoelectron microscopy (*271*) and cryoelectron tomography (*272*, *273*) have notably revealed that the ultrastructure of bacterial and archaeal cells is far more complex and diverse than assumed previously (*270*, *274–276*). Microorganisms are now known to have a wide variety of intracellular organelles (*275*), as well as other intracellular compartments of unknown function including nanospheres and both intracellular and

periplasmic vesicles (*274*). Further, bacterial and archaeal cells often include various types of intracellular filaments, bundles, arrays, and tubes in addition to varied cell appendages (*274*). The extent to which the cytological features of certain bacteria and archaea, such as *Ca.* P. syntrophicum (*121*), are related to one another and to those of eukaryotes, remains for now largely unknown considering that genes and proteins involved in their formation have not been identified in many cases. Current advances in the computational prediction of the structure of individual proteins (*277*, *278*) and both the composition and structure of protein complexes (*277*, *279*) have the potential to accelerate the identification of genes involved in protein complexes forming cytological features. Indeed, the accuracy of the protein structures predicted by the neural-network models AlphaFold2 (*278*) and RoseTTA fold (*277*) rivals that of experimentally determined structures (*277*, *280*). Predicted protein structures can help interpreting Coulomb potential maps obtained by cryoelectron microscopy and cellular cryoelectron tomography for the experimental determination of protein structures (*281*). Furthermore, the development of standards to adequately evaluate the fit of computationally predicted protein models to the Coulomb potential maps of protein complexes may allow to refine protein complex structures and identify genes coding for protein complex components (*282*). We envision that progress in the computational predictions of protein structures may also allow for the identification of proteins, which share similar folds but little to no amino acid sequence similarity to known components of well-characterized cellular features. Once candidate protein components of a cellular feature of interest have been identified by, for instance, immunogold labeling (*283*), the localization, dynamics, and function of the proteins, and corresponding cytological features can be investigated using antibodies conjugated with fluorescent labels and superresolution microscopy (*284*, *285*) as performed, for example, for the analysis of the cytokinesis machinery of bacteria (*286*) and archaea (*287*). Altogether, these protein structure-based approaches combined with high-end microscopy now allow us to bridge the gap between bioinformatic analyses and cell biology and to reconstruct major steps in the evolution of cellular complexity.

GENOTYPE–PHENOTYPE RELATIONSHIP

Moving from the reconstruction of the evolutionary history of life to understanding the evolutionary trajectories taken by life forms through time requires clarifying their evolvability (*288–290*). This includes elucidating the physical constraints on the phenotypes that organisms or their cellular components may take (*291–294*) but also identifying features of biological systems opening opportunities for the emergence of phenotypic variation, innovation, and diversification (*295*). This emphasizes the need to study fundamental attributes of microbial cells including for example, trade-offs (*296*, *297*), allometric scaling laws (*298–300*) and robustness (*301–303*) and their respective underlying causes at the molecular level. Progress in this research area will allow for a better understanding of the relation between genotype and phenotype (i.e., genotype-phenotype map (*304–306*)) thereby clarifying the landscape of possible genetic changes. Advances in high-throughput phenotyping and genotyping, targeted genome editing, and single cell approaches (*307–315*), evolutionary synthetic biology (*316–318*), and experimental evolution (*319*), are currently driving progress in the exploration of the genotype–phenotype map. Yet, conceptual, and theoretical developments need to follow technological advances to derive the

principles determining the evolution of (micro)organisms. Although such studies are typically conducted on model organisms, a focus on microbial groups placed near key evolutionary branching points would be beneficial for understanding major transitions in the early evolution of life on Earth. This emphasizes the need to isolate and develop laboratory cultivation systems to study members of these microbial groups, most of which remain currently uncultivated (*320*).

CONCLUSION

The TOL is a constantly changing and evolving concept in evolutionary biology, which has helped to depict the vast biodiversity on Earth, including both vertical and horizontal relations of organisms as well as connections to MGEs including viruses. Of course, it will always constitute a simplified illustration of the diversification of life on Earth and can only account for the evolutionary path of extant organisms even though extinct organisms may have contributed to the genetic repertoire of extant genomes. For example, all organisms today are derived from LUCA, yet the early diversification of LUCA was likely shaped by gene influx from now extinct organisms living at the time of LUCA.

Nevertheless, the TOL provides a useful concept for describing and classifying the diversity of organismal life on Earth today (*25*, *26*) and for improving our understanding of events leading to major evolutionary changes that have dramatically impacted our biosphere. The continuous improvement of analytical, experimental and computational approaches to the study of life's biodiversity and integration of geological records will further improve our insights into the evolutionary past and allow linking diversification to Earth history. Further, this will help to refine our understanding of evolutionary principles underlying biodiversification, which is crucial for predicting evolution and may help efforts to preserve biodiversity in an ever-changing world.

ACKNOWLEDGMENTS

A.S. was supported by the European Research Council (ERC STG ASymbEL: 947317), the Swedish Research Council (VR starting grant 2016-03559 to A.S.) and the NWO-I foundation of the Netherlands Organisation for Scientific Research (WISE fellowship). C.W.S. was supported by the Swedish Research Council (VR starting grant 2020-05071). We apologize to colleagues whose work we were unable to cite due to length constraints.

Author Contributions

A.S. conceptualized and all authors have contributed to the writing of this review.

Author notes

† Tara A. Mahendrarajah, Pierre Offre and Courtney W. Stairs contributed equally to this work.

REFERENCES

1. C. R. Woese, G. E. Fox, Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5088–5090 (1977).

2. C. R. Woese, O. Kandler, M. L. Wheelis, Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 4576–4579 (1990).

3. M. C. Weiss, F. L. Sousa, N. Mrnjavac, S. Neukirchen, M. Roettger, S. Nelson-Sathi, W. F. Martin, The physiology and habitat of the last universal common ancestor. *Nat Microbiol* **1**, 16116 (2016).

4. T. Dagan, Y. Artzy-Randrup, W. Martin, Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 10039–10044 (2008).

5. P. Puigbò, Y. I. Wolf, E. V. Koonin, Search for a "Tree of Life" in the thicket of the phylogenetic forest. *J. Biol.* **8**, 59 (2009).

6. C. Blais, J. M. Archibald, The past, present and future of the tree of life. *Curr. Biol.* **31**, R314–R321 (2021).

7. L. Guy, J. H. Saw, T. J. G. Ettema, The archaeal legacy of eukaryotes: a phylogenomic perspective. *Cold Spring Harb. Perspect. Biol.* **6**, a016022 (2014).

8. E. V. Koonin, N. Yutin, The dispersed archaeal eukaryome and the complex archaeal ancestor of eukaryotes. *Cold Spring Harb. Perspect. Biol.* **6**, a016188 (2014).

9. P. López-García, D. Moreira, The Syntrophy hypothesis for the origin of eukaryotes revisited. *Nat Microbiol* **5**, 655–667 (2020).

10. W. F. Martin, S. Garg, V. Zimorski, Endosymbiotic theories for eukaryote origin. (2015). https://doi.org/10.1098/rstb.2014.0330.

11. L. Eme, A. Spang, J. Lombard, C. W. Stairs, T. J. G. Ettema, Archaea and the origin of eukaryotes. *Nat. Rev. Microbiol.* **15**, 711–723 (2017).

12. T. A. Williams, P. G. Foster, C. J. Cox, T. M. Embley, An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* **504**, 231–236 (2013).

13. T. A. Williams, C. J. Cox, P. G. Foster, G. J. Szöllősi, T. M. Embley, Phylogenomics provides robust support for a two-domains tree of life. *Nat Ecol Evol* **4**, 138–147 (2020).

14. E. V. Koonin, P. Starokadomskyy, Are viruses alive? The replicator paradigm sheds decisive light on an old but misguided question. *Stud. Hist. Philos. Biol. Biomed. Sci.* **59**, 125–134 (2016).

15. D. Moreira, P. López-García, Ten reasons to exclude viruses from the tree of life. *Nat. Rev. Microbiol.* **7**, 306–311 (2009).

16. O. Popa, T. Dagan, Trends and barriers to lateral gene transfer in prokaryotes. *Curr. Opin. Microbiol.* **14**, 615–623 (2011).

17. E. V. Koonin, Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions. *F1000Res.* **5**, 1805 (2016).

18. W. F. Doolittle, E. Bapteste, Pattern pluralism and the Tree of Life hypothesis. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 2043–2049 (2007).

19. E. V. Koonin, The turbulent network dynamics of microbial evolution and the statistical tree of life. *J. Mol. Evol.* **80**, 244–250 (2015).

20. L. A. Hug, B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D. A. Relman, K. M. Finstad, R. Amundson, B. C. Thomas, J. F. Banfield, A new view of the tree of life. *Nat Microbiol* **1**, 16048 (2016).

21. M. Krupovic, V. V. Dolja, E. V. Koonin, The LUCA and its complex virome. *Nat. Rev. Microbiol.* **18**, 661–670 (2020).

22. P. S. Adam, G. Borrel, C. Brochier-Armanet, S. Gribaldo, The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. *ISME J.* **11**, 2407–2425 (2017).

23. A. Spang, E. F. Caceres, T. J. G. Ettema, Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. *Science* **357**, eaaf3883 (2017).

24. C. J. Castelle, J. F. Banfield, Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. *Cell* **172**, 1181–1197 (2018).

25. D. H. Parks, M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil, P. Hugenholtz, A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).

26. C. Rinke, M. Chuvochina, A. J. Mussig, P.-A. Chaumeil, A. A. Davín, D. W. Waite, W. B. Whitman, D. H. Parks, P. Hugenholtz, A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nat Microbiol* **6**, 946–959 (2021).

27. F. Burki, A. J. Roger, M. W. Brown, A. G. B. Simpson, The new tree of eukaryotes. *Trends Ecol. Evol.* **35**, 43–55 (2020).

28. D. Paez-Espino, E. A. Eloe-Fadrosh, G. A. Pavlopoulos, A. D. Thomas, M. Huntemann, N. Mikhailova, E. Rubin, N. N. Ivanova, N. C. Kyrpides, Uncovering earth's virome. *Nature* **536**, 425–430 (2016).

29. F. Martinez-Hernandez, O. Fornas, M. Lluesma Gomez, B. Bolduc, M. J. de la Cruz Peña, J. M. Martínez, J. Anton, J. M. Gasol, R. Rosselli, F. Rodriguez-Valera, M. B. Sullivan, S. G. Acinas, M. Martinez-Garcia, Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nat. Commun.* **8**, 15892 (2017).

30. A. C. Gregory, A. A. Zayed, N. Conceição-Neto, B. Temperton, B. Bolduc, A. Alberti, M. Ardyna, K. Arkhipova, M. Carmichael, C. Cruaud, C. Dimier, G. Domínguez-Huerta, J. Ferland, S. Kandels, Y. Liu, C. Marec, S. Pesant, M. Picheral, S. Pisarev, J. Poulain, J.-É. Tremblay, D. Vik, Tara Oceans Coordinators, M. Babin, C. Bowler, A. I. Culley, C. de Vargas, B. E. Dutilh, D. Iudicone, L. Karp-Boss, S. Roux, S. Sunagawa, P. Wincker, M. B. Sullivan, Marine DNA viral macro- and microdiversity from pole to pole. *Cell* **177**, 1109-1123.e14 (2019).

31. J. Beaulaurier, E. Luo, J. M. Eppley, P. D. Uyl, X. Dai, A. Burger, D. J. Turner, M. Pendelton, S. Juul, E. Harrington, E. F. DeLong, Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities. *Genome Res.* **30**, 437–446 (2020).

32. M. Moniruzzaman, C. A. Martinez-Gutierrez, A. R. Weinheimer, F. O. Aylward, Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nat. Commun.* **11**, 1710 (2020).

33. C. M. Bellas, R. Sommaruga, Polinton-like viruses are abundant in aquatic ecosystems. *Microbiome* **9**, 13 (2021).

34. R. C. Edgar, B. Taylor, V. Lin, T. Altman, P. Barbera, D. Meleshko, D. Lohr, G. Novakovsky, B. Buchfink, B. Al-Shayeb, J. F. Banfield, M. de la Peña, A. Korobeynikov, R. Chikhi, A. Babaian, Petabase-scale sequence alignment catalyses viral discovery. *Nature* **602**, 142–147 (2022).

35. S. Roux, S. J. Hallam, T. Woyke, M. B. Sullivan, Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *Elife* **4** (2015).

36. M. Džunková, S. J. Low, J. N. Daly, L. Deng, C. Rinke, P. Hugenholtz, Defining the human gut host-phage network through single-cell viral tagging. *Nat. Microbiol.* **4**, 2192–2203 (2019).

37. J. K. Jarett, M. Džunková, F. Schulz, S. Roux, D. Paez-Espino, E. Eloe-Fadrosh, S. P. Jungbluth, N. Ivanova, J. R. Spear, S. A. Carr, C. B. Trivedi, F. A. Corsetti, H. A. Johnson, E. Becraft, N. Kyrpides, R. Stepanauskas, T. Woyke, Insights into the dynamics between viruses and their hosts in a hot spring microbial mat. *ISME J.* **14**, 2527–2541 (2020).

38. E. G. Sakowski, K. Arora-Williams, F. Tian, A. A. Zayed, O. Zablocki, M. B. Sullivan, S. P. Preheim, Interaction dynamics and virus-host range for estuarine actinophages captured by epicPCR. *Nat. Microbiol.* **6**, 630–642 (2021).

39. M. Krupovic, V. V. Dolja, E. V. Koonin, Origin of viruses: primordial replicators recruiting capsids from hosts. *Nat. Rev. Microbiol.* **17**, 449–458 (2019).

40. B. Schoepp-Cothenet, R. van Lis, A. Atteia, F. Baymann, L. Capowiez, A.-L. Ducluzeau, S. Duval, F. ten Brink, M. J. Russell, W. Nitschke, On the universal core of bioenergetics. *Biochim. Biophys. Acta* **1827**, 79–93 (2013).

41. F. L. Sousa, T. Thiergart, G. Landan, S. Nelson-Sathi, I. A. C. Pereira, J. F. Allen, N. Lane, W. F. Martin, Early bioenergetic evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368**, 20130088 (2013).

42. V. Sojo, A. Pomiankowski, N. Lane, A bioenergetic basis for membrane divergence in archaea and bacteria. *PLoS Biol.* **12**, e1001926 (2014).

43. M. J. Russell, W. Nitschke, Methane: Fuel or Exhaust at the Emergence of Life? *Astrobiology* **17**, 1053–1066 (2017).

44. N. Iwabe, K. Kuma, M. Hasegawa, S. Osawa, T. Miyata, Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 9355–9359 (1989).

45. J. R. Brown, W. F. Doolittle, Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 2441–2445 (1995).

46. O. Zhaxybayeva, P. Lapierre, J. P. Gogarten, Ancient gene duplications and the root(s) of the tree of life. *Protoplasma* **227**, 53–64 (2005).

47. A. R. Weinheimer, F. O. Aylward, A distinct lineage of Caudovirales that encodes a deeply branching multi-subunit RNA polymerase. *Nat. Commun.* **11**, 4506 (2020).

48. T. Dagan, M. Roettger, D. Bryant, W. Martin, Genome networks root the tree of life between prokaryotic domains. *Genome Biol. Evol.* **2**, 379–392 (2010).

49. T. Cavalier-Smith, Rooting the tree of life by transition analyses. *Biol. Direct* **1**, 19 (2006).

50. J. A. Lake, R. G. Skophammer, C. W. Herbold, J. A. Servin, Genome beginnings: rooting the tree of life. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 2177–2185 (2009).

51. R. Gouy, D. Baurain, H. Philippe, Rooting the tree of life: the phylogenetic jury is still out. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140329 (2015).

52. Q. Zhu, U. Mai, W. Pfeiffer, S. Janssen, F. Asnicar, J. G. Sanders, P. Belda-Ferre, G. A. Al-Ghalith, E. Kopylova, D. McDonald, T. Kosciolek, J. B. Yin, S. Huang, N. Salam, J.-Y. Jiao, Z. Wu, Z. Z. Xu, K. Cantrell, Y. Yang, E. Sayyari, M. Rabiee, J. T. Morton, S. Podell, D. Knights, W.-J. Li, C. Huttenhower, N. Segata, L. Smarr, S. Mirarab, R. Knight, Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477 (2019).

53. C. A. Martinez-Gutierrez, F. O. Aylward, Phylogenetic Signal, Congruence, and Uncertainty across Bacteria and Archaea. *Mol. Biol. Evol.* **38**, 5514–5527 (2021).

54. E. R. R. Moody, T. A. Mahendrarajah, N. Dombrowski, J. W. Clark, C. Petitjean, P. Offre, G. J. Szöllősi, A. Spang, T. A. Williams, An estimate of the deepest branches of the tree of life from ancient vertically evolving genes. *Elife* **11** (2022).

55. G. J. Szöllosi, B. Boussau, S. S. Abby, E. Tannier, V. Daubin, Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17513–17518 (2012).

56. L. A. David, E. J. Alm, Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* **469**, 93–96 (2011).

57. G. J. Szöllõsi, W. Rosikiewicz, B. Boussau, E. Tannier, V. Daubin, Efficient exploration of the space of reconciled gene trees. *Syst. Biol.* **62**, 901–912 (2013).

58. C. Rinke, P. Schwientek, A. Sczyrba, N. N. Ivanova, I. J. Anderson, J.-F. Cheng, A. Darling, S. Malfatti, B. K. Swan, E. A. Gies, J. A. Dodsworth, B. P. Hedlund, G. Tsiamis, S. M. Sievert, W.-T. Liu, J. A. Eisen, S. J. Hallam, N. C. Kyrpides, R. Stepanauskas, E. M. Rubin, P. Hugenholtz, T. Woyke, Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).

59. C. J. Castelle, K. C. Wrighton, B. C. Thomas, L. A. Hug, C. T. Brown, M. J. Wilkins, K. R. Frischkorn, S. G. Tringe, A. Singh, L. M. Markillie, R. C. Taylor, K. H. Williams, J. F. Banfield, Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr. Biol.* **25**, 690–701 (2015).

60. C. T. Brown, L. A. Hug, B. C. Thomas, I. Sharon, C. J. Castelle, A. Singh, M. J. Wilkins, K. C. Wrighton, K. H. Williams, J. F. Banfield, Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).

61. T. A. Williams, G. J. Szöllősi, A. Spang, P. G. Foster, S. E. Heaps, B. Boussau, T. J. G. Ettema, T. M. Embley, Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E4602–E4611 (2017).

62. C. J. Castelle, C. T. Brown, K. Anantharaman, A. J. Probst, R. H. Huang, J. F. Banfield, Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat. Rev. Microbiol.* **16**, 629–645 (2018).

63. N. Taib, D. Megrian, J. Witwinowski, P. Adam, D. Poppleton, G. Borrel, C. Beloin, S. Gribaldo, Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. *Nat Ecol Evol* **4**, 1661–1672 (2020).

64. G. A. Coleman, A. A. Davín, T. A. Mahendrarajah, L. L. Szánthó, A. Spang, P. Hugenholtz, G. J. Szöllősi, T. A. Williams, A rooted phylogeny resolves early bacterial evolution. *Science* **372** (2021).

65. J. C. Xavier, R. E. Gerhards, J. L. E. Wimmer, J. Brueckner, F. D. K. Tria, W. F. Martin, The metabolic network of the last bacterial common ancestor. (2021). https://doi.org/10.1038/s42003-021-01918-4.

66. H. Huber, M. J. Hohn, R. Rachel, T. Fuchs, V. C. Wimmer, K. O. Stetter, A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* **417**, 63–67 (2002).

67. M. Podar, K. S. Makarova, D. E. Graham, Y. I. Wolf, E. V. Koonin, A.-L. Reysenbach, Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone National Park. *Biol. Direct* **8**, 9 (2013).

68. J. H. Munson-McGee, E. K. Field, M. Bateson, C. Rooney, R. Stepanauskas, M. J. Young, Nanoarchaeota, their Sulfolobales host, and Nanoarchaeota virus distribution across Yellowstone National Park Hot Springs. *Appl. Environ. Microbiol.* **81**, 7860–7868 (2015).

69. L. Wurch, R. J. Giannone, B. S. Belisle, C. Swift, S. Utturkar, R. L. Hettich, A.-L. Reysenbach, M. Podar, Genomics-informed isolation and characterization of a symbiotic Nanoarchaeota system from a terrestrial geothermal environment. *Nat. Commun.* **7**, 12115 (2016).

70. O. V. Golyshina, S. V. Toshchakov, K. S. Makarova, S. N. Gavrilov, A. A. Korzhenkov, V. La Cono, E. Arcadi, T. Y. Nechitaylo, M. Ferrer, I. V. Kublanov, Y. I. Wolf, M. M. Yakimov, P. N. Golyshin, 'ARMAN' archaea depend on association with euryarchaeal host in culture and in situ. *Nat. Commun.* **8** (2017).

71. S. Krause, A. Bremges, P. C. Münch, A. C. McHardy, J. Gescher, Characterisation of a stable laboratory co-culture of acidophilic nanoorganisms. *Sci. Rep.* **7**, 3289 (2017).

72. J. N. Hamm, S. Erdmann, E. A. Eloe-Fadrosh, A. Angeloni, L. Zhong, C. Brownlee, T. J. Williams, K. Barton, S. Carswell, M. A. Smith, S. Brazendale, A. M. Hancock, M. A. Allen, M. J. Raftery, R. Cavicchioli, Unexpected host dependency of Antarctic Nanohaloarchaeota. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 14661–14670 (2019).

73. E. St. John, Y. Liu, M. Podar, M. B. Stott, J. Meneghin, Z. Chen, K. Lagutin, K. Mitchell, A.-L. Reysenbach, A new symbiotic nanoarchaeote (Candidatus Nanoclepta minutus) and its host (Zestosphaera tikiterensis gen. nov., sp. nov.) from a New Zealand hot spring. *Syst. Appl. Microbiol.* **42**, 94–106 (2019).

74. V. La Cono, E. Messina, M. Rohde, E. Arcadi, S. Ciordia, F. Crisafi, R. Denaro, M. Ferrer, L. Giuliano, P. N. Golyshin, O. V. Golyshina, J. E. Hallsworth, G. La Spada, M. C. Mena, A. Y. Merkel, M. A. Shevchenko, F. Smedile, D. Y. Sorokin, S. V. Toshchakov, M. M. Yakimov, Symbiosis between nanohaloarchaeon and haloarchaeon is based on utilization of different polysaccharides. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 20223–20234 (2020).

75. H. D. Sakai, N. Nur, S. Kato, M. Yuki, M. Shimizu, T. Itoh, M. Ohkuma, A. Suwanto, N. Kurosawa, Insight into the symbiotic lifestyle of DPANN archaea revealed by cultivation and genome analyses. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2115449119 (2022).

76. X. He, J. S. McLean, A. Edlund, S. Yooseph, A. P. Hall, S.-Y. Liu, P. C. Dorrestein, E. Esquenazi, R. C. Hunter, G. Cheng, K. E. Nelson, R. Lux, W. Shi, Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 244–249 (2015).

77. B. Bor, J. S. McLean, K. R. Foster, L. Cen, T. T. To, A. Serrato-Guillen, F. E. Dewhirst, W. Shi, X. He, Rapid evolution of decreased host susceptibility drives a stable relationship between ultrasmall parasite TM7x and its bacterial host. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 12277–12282 (2018).

78. D. Moreira, Y. Zivanovic, A. I. López-Archilla, M. Iniesto, P. López-García, Reductive evolution and unique predatory mode in the CPR bacterium Vampirococcus lugosii. *Nat. Commun.* **12**, 2454 (2021).

79. U. Jahn, M. Gallenberger, W. Paper, B. Junglas, W. Eisenreich, K. O. Stetter, R. Rachel, H. Huber, Nanoarchaeum equitans and Ignicoccus hospitalis: new insights into a unique, intimate association of two archaea. *J. Bacteriol.* **190**, 1743–1750 (2008).

80. A. J. Probst, T. Weinmaier, K. Raymann, A. Perras, J. B. Emerson, T. Rattei, G. Wanner, A. Klingl, I. A. Berg, M. Yoshinaga, B. Viehweger, K.-U. Hinrichs, B. C. Thomas, S. Meck, A. K. Auerbach, M. Heise, A. Schintlmeister, M. Schmid, M. Wagner, S. Gribaldo, J. F. Banfield, C. Moissl-Eichinger, Biology of a widespread uncultivated archaeon that contributes to carbon fixation in the subsurface. *Nat. Commun.* **5**, 5497 (2014).

81. J. P. Beam, E. D. Becraft, J. M. Brown, F. Schulz, J. K. Jarett, O. Bezuidt, N. J. Poulton, K. Clark, P. F. Dunfield, N. V. Ravin, J. R. Spear, B. P. Hedlund, K. A. Kormas, S. M. Sievert, M. S. Elshahed, H. A. Barton, M. B. Stott, J. A. Eisen, D. P. Moser, T. C. Onstott, T. Woyke, R. Stepanauskas, Ancestral absence of electron transport chains in Patescibacteria and DPANN. *Front. Microbiol.* **11**, 1848 (2020).

82. N. H. Youssef, C. Rinke, R. Stepanauskas, I. Farag, T. Woyke, M. S. Elshahed, Insights into the metabolism, lifestyle and putative evolutionary history of the novel archaeal phylum "Diapherotrites." *ISME J.* **9**, 447–460 (2015).

83. R. Méheust, D. Burstein, C. J. Castelle, J. F. Banfield, The distinction of CPR bacteria from other bacteria based on protein family content. *Nat. Commun.* **10**, 4173 (2019).

84. N. Dombrowski, J.-H. Lee, T. A. Williams, P. Offre, A. Spang, Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS Microbiol. Lett.* **366** (2019).

85. C. Brochier-Armanet, P. Forterre, S. Gribaldo, Phylogeny and evolution of the Archaea: one hundred genomes later. *Curr. Opin. Microbiol.* **14**, 274–281 (2011).

86. C. Petitjean, P. Deschamps, P. López-García, D. Moreira, Rooting the domain archaea by phylogenomic analysis supports the foundation of the new kingdom Proteoarchaeota. *Genome Biol. Evol.* **7**, 191–204 (2014).

87. K. Raymann, P. Forterre, C. Brochier-Armanet, S. Gribaldo, Global phylogenomic analysis disentangles the complex evolutionary history of DNA replication in archaea. *Genome Biol. Evol.* **6**, 192–212 (2014).

88. M. Aouad, N. Taib, A. Oudart, M. Lecocq, M. Gouy, C. Brochier-Armanet, Extreme halophilic archaea derive from two distinct methanogen Class II lineages. *Mol. Phylogenet. Evol.* **127**, 46–54 (2018).

89. Y. Feng, U. Neri, S. Gosselin, A. S. Louyakis, R. T. Papke, U. Gophna, J. P. Gogarten, The Evolutionary Origins of Extreme Halophilic Archaeal Lineages. *Genome Biol. Evol.* **13** (2021).

90. J. Bergsten, A review of long-branch attraction. *Cladistics* **21**, 163–193 (2005).

91. H. Philippe, Y. Zhou, H. Brinkmann, N. Rodrigue, F. Delsuc, Heterotachy and long-branch attraction in phylogenetics. *BMC Evol. Biol.* **5**, 50 (2005).

92. N. A. Moran, Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 2873–2878 (1996).

93. B. Rodriguez-Brito, F. Rohwer, R. A. Edwards, An application of statistics to comparative metagenomics. *BMC Bioinformatics* **7**, 162 (2006).

94. N. Dombrowski, T. A. Williams, J. Sun, B. J. Woodcroft, J.-H. Lee, B. Q. Minh, C. Rinke, A. Spang, Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. *Nat. Commun.* **11**, 3939 (2020).

95. C. J. Castelle, R. Méheust, A. L. Jaffe, K. Seitz, X. Gong, B. J. Baker, J. F. Banfield, Protein family content uncovers lineage relationships and bacterial pathway maintenance mechanisms in DPANN Archaea. *Front. Microbiol.* **12**, 660052 (2021).

96. M. Aouad, J.-P. Flandrois, F. Jauffrit, M. Gouy, S. Gribaldo, C. Brochier-Armanet, A divide-and-conquer phylogenomic approach based on character supermatrices resolves early steps in the evolution of the Archaea. *BMC Ecol. Evol.* **22**, 1 (2022).

97. P. López-García, D. Moreira, Open questions on the origin of eukaryotes. *Trends Ecol. Evol.* **30**, 697–708 (2015).

98. E. V. Koonin, Origin of eukaryotes from within archaea, archaeal eukaryome and bursts of gene gain: eukaryogenesis just made easier? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140333 (2015).

99. A. J. Roger, S. A. Muñoz-Gómez, R. Kamikawa, The Origin and Diversification of Mitochondria. *Curr. Biol.* **27**, R1177–R1192 (2017).

100. A. Spang, J. H. Saw, S. L. Jørgensen, K. Zaremba-Niedzwiedzka, J. Martijn, A. E. Lind, R. van Eijk, C. Schleper, L. Guy, T. J. G. Ettema, Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).

101. K. Zaremba-Niedzwiedzka, E. F. Caceres, J. H. Saw, D. Bäckström, L. Juzokaite, E. Vancaester, K. W. Seitz, K. Anantharaman, P. Starnawski, K. U. Kjeldsen, M. B. Stott, T. Nunoura, J. F. Banfield, A. Schramm, B. J. Baker, A. Spang, T. J. G. Ettema, Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).

102. K. W. Seitz, C. S. Lazar, K.-U. Hinrichs, A. P. Teske, B. J. Baker, Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. *ISME J.* **10**, 1696–1705 (2016).

103. K. W. Seitz, N. Dombrowski, L. Eme, A. Spang, J. Lombard, J. R. Sieber, A. P. Teske, T. J. G. Ettema, B. J. Baker, Asgard archaea capable of anaerobic hydrocarbon cycling. *Nat. Commun.* **10**, 1822 (2019).

104. M. Cai, Y. Liu, X. Yin, Z. Zhou, M. W. Friedrich, T. Richter-Heitmann, R. Nimzyk, A. Kulkarni, X. Wang, W. Li, J. Pan, Y. Yang, J.-D. Gu, M. Li, Diverse Asgard archaea including the novel phylum Gerdarchaeota participate in organic matter degradation. *Sci. China Life Sci.* **63**, 886–897 (2020).

105. Y. Liu, K. S. Makarova, W.-C. Huang, Y. I. Wolf, A. N. Nikolskaya, X. Zhang, M. Cai, C.-J. Zhang, W. Xu, Z. Luo, L. Cheng, E. V. Koonin, M. Li, Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature* **593**, 553–557 (2021).

106. J.-W. Zhang, H.-P. Dong, L.-J. Hou, Y. Liu, Y.-F. Ou, Y.-L. Zheng, P. Han, X. Liang, G.-Y. Yin, D.-M. Wu, M. Liu, M. Li, Newly discovered Asgard archaea Hermodarchaeota potentially degrade alkanes and aromatics via alkyl/benzyl-succinate synthase and benzoyl-CoA pathway. *ISME J.* **15**, 1826–1843 (2021).

107. F. Wu, D. R. Speth, A. Philosof, A. Crémière, A. Narayanan, R. A. Barco, S. A. Connon, J. P. Amend, I. A. Antoshechkin, V. J. Orphan, Unique mobile elements and scalable gene flow at the prokaryote–eukaryote boundary revealed by circularized Asgard archaea genomes. *Nature Microbiology* **7**, 200–212 (2022).

108. I. F. Farag, R. Zhao, J. F. Biddle, "Sifarchaeota," a Novel Asgard Phylum from Costa Rican Sediment Capable of Polysaccharide Degradation and Anaerobic Methylotrophy. *Appl. Environ. Microbiol.* **87** (2021).

109. H. Hartman, A. Fedorov, The origin of the eukaryotic cell: a genomic investigation. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 1420–1425 (2002).

110. C. Akıl, R. C. Robinson, Genomes of Asgard archaea encode profilins that regulate actin. *Nature* **562**, 439–443 (2018).

111. C. Akıl, L. T. Tran, M. Orhant-Prioux, Y. Baskaran, E. Manser, L. Blanchoin, R. C. Robinson, Insights into the evolution of regulated actin dynamics via characterization of primitive gelsolin/cofilin proteins from Asgard archaea. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 19904–19913 (2020).

112. S. Survery, F. Hurtig, S. R. Haq, J. Eriksson, L. Guy, K. J. Rosengren, A.-C. Lindås, C. N. Chi, Heimdallarchaea encodes profilin with eukaryotic-like actin regulation and polyproline binding. *Commun. Biol.* **4**, 1024 (2021).

113. R. M. Bowers, The Genome Standards Consortium, N. C. Kyrpides, R. Stepanauskas, M. Harmon-Smith, D. Doud, T. B. K. Reddy, F. Schulz, J. Jarett, A. R. Rivers, E. A. Eloe-Fadrosh, S. G. Tringe, N. N. Ivanova, A. Copeland, A. Clum, E. D. Becraft, R. R. Malmstrom, B. Birren, M. Podar, P. Bork, G. M. Weinstock, G. M. Garrity, J. A. Dodsworth, S. Yooseph, G. Sutton, F. O. Glöckner, J. A. Gilbert, W. C. Nelson, S. J. Hallam, S. P. Jungbluth, T. J. G. Ettema, S. Tighe, K. T. Konstantinidis, W.-T. Liu, B. J. Baker, T. Rattei, J. A. Eisen, B. Hedlund, K. D. McMahon, N. Fierer, R. Knight, R. Finn, G. Cochrane, I. Karsch-Mizrachi, G. W. Tyson, C. Rinke, A. Lapidus, F. Meyer, P. Yilmaz, D. H. Parks, A. Murat Eren, L. Schriml, J. F. Banfield, P. Hugenholtz, T. Woyke, Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).

114. V. Da Cunha, M. Gaia, D. Gadelle, A. Nasir, P. Forterre, Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS Genet.* **13**, e1006810 (2017).

115. V. Da Cunha, M. Gaia, A. Nasir, P. Forterre, Asgard archaea do not close the debate about the universal tree of life topology, *PLoS genetics*. **14** (2018)p. e1007215.

116. S. G. Garg, N. Kapust, W. Lin, M. Knopp, F. D. K. Tria, S. Nelson-Sathi, S. B. Gould, L. Fan, R. Zhu, C. Zhang, W. F. Martin, Anomalous Phylogenetic Behavior of Ribosomal Proteins in Metagenome-Assembled Asgard Archaea. *Genome Biol. Evol.* **13** (2021).

117. A. Spang, L. Eme, J. H. Saw, E. F. Caceres, K. Zaremba-Niedzwiedzka, J. Lombard, L. Guy, T. J. G. Ettema, Asgard archaea are the closest prokaryotic relatives of eukaryotes, *PLoS genetics*. **14** (2018)p. e1007080.

118. L. Manoharan, J. A. Kozlowski, R. W. Murdoch, F. E. Löffler, F. L. Sousa, C. Schleper, Metagenomes from coastal marine sediments give insights into the ecological role and cellular features of Loki- and Thorarchaeota. *MBio* **10** (2019).

119. R. Chen, H. L. Wong, G. S. Kindler, F. I. MacLeod, N. Benaud, B. C. Ferrari, B. P. Burns, Discovery of an abundance of biosynthetic gene clusters in shark bay microbial mats. *Front. Microbiol.* **11** (2020).

120. I. F. Farag, J. F. Biddle, R. Zhao, A. J. Martino, C. H. House, R. I. León-Zayas, Metabolic potentials of archaeal lineages resolved from metagenomes of deep Costa Rica sediments. *ISME J.* **14**, 1345–1358 (2020).

121. H. Imachi, M. K. Nobu, N. Nakahara, Y. Morono, M. Ogawara, Y. Takaki, Y. Takano, K. Uematsu, T. Ikuta, M. Ito, Y. Matsui, M. Miyazaki, K. Murata, Y. Saito, S. Sakai, C. Song, E. Tasumi, Y. Yamanaka, T. Yamaguchi, Y. Kamagata, H. Tamaki, K. Takai, Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**, 519–525 (2020).

122. B. Avcı, J. Brandt, D. Nachmias, N. Elia, M. Albertsen, T. J. G. Ettema, A. Schramm, K. U. Kjeldsen, Spatial separation of ribosomes and DNA in Asgard archaeal cells. *ISME J.* **16**, 606–610 (2022).

123. A. Spang, C. W. Stairs, N. Dombrowski, L. Eme, J. Lombard, E. F. Caceres, C. Greening, B. J. Baker, T. J. G. Ettema, Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. *Nat Microbiol* **4**, 1138–1148 (2019).

124. W. Martin, M. Müller, The hydrogen hypothesis for the first eukaryote. *Nature* **392**, 37–41 (1998).

125. D. Moreira, P. Lopez-Garcia, Symbiosis between methanogenic archaea and delta-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. *J. Mol. Evol.* **47**, 517–530 (1998).

126. J. Martijn, J. Vosseberg, L. Guy, P. Offre, T. J. G. Ettema, Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* **557**, 101–105 (2018).

127. L. Fan, D. Wu, V. Goremykin, J. Xiao, Y. Xu, S. Garg, C. Zhang, W. F. Martin, R. Zhu, Phylogenetic analyses with systematic taxon sampling show that mitochondria branch within Alphaproteobacteria. *Nat Ecol Evol* **4**, 1213–1219 (2020).

128. S. A. Muñoz-Gómez, E. Susko, K. Williamson, L. Eme, C. H. Slamovits, D. Moreira, P. López-García, A. J. Roger, Site-and-branch-heterogeneous analyses of an expanded dataset favour mitochondria as sister to known Alphaproteobacteria. *Nat Ecol Evol* **6**, 253–262 (2022).

129. D. A. Baum, B. Baum, An inside-out origin for the eukaryotic cell. *BMC Biol.* **12**, 76 (2014).

130. A. M. Poole, S. Gribaldo, Eukaryotic origins: How and when was the mitochondrion acquired? *Cold Spring Harb. Perspect. Biol.* **6**, a015990 (2014).

131. S. B. Gould, S. G. Garg, W. F. Martin, Bacterial vesicle secretion and the evolutionary origin of the eukaryotic endomembrane system. *Trends Microbiol.* **24**, 525–534 (2016).

132. A. A. Pittis, T. Gabaldón, Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature* **531**, 101–104 (2016).

133. F. D. K. Tria, J. Brueckner, J. Skejo, J. C. Xavier, N. Kapust, M. Knopp, J. L. E. Wimmer, F. S. P. Nagies, V. Zimorski, S. B. Gould, S. G. Garg, W. F. Martin, Gene duplications trace mitochondria to the onset of eukaryote complexity. *Genome Biol. Evol.* **13** (2021).

134. J. Vosseberg, J. J. E. van Hooff, M. Marcet-Houben, A. van Vlimmeren, L. M. van Wijk, T. Gabaldón, B. Snel, Timing the origin of eukaryotic cellular complexity with ancient duplications. *Nat Ecol Evol* **5**, 92–100 (2021).

135. E. V. Koonin, The origin and early evolution of eukaryotes in the light of phylogenomics. *Genome Biol.* **11**, 209 (2010).

136. N. C. Rochette, C. Brochier-Armanet, M. Gouy, Phylogenomic test of the hypotheses for the evolutionary origin of eukaryotes. *Mol. Biol. Evol.* **31**, 832–845 (2014).

137. C. Santana-Molina, E. Rivas-Marin, A. M. Rojas, D. P. Devos, Origin and Evolution of Polycyclic Triterpene Synthesis. *Mol. Biol. Evol.* **37**, 1925–1941 (2020).

138. C. W. Stairs, J. E. Dharamshi, D. Tamarit, L. Eme, S. L. Jørgensen, A. Spang, T. J. G. Ettema, Chlamydial contribution to anaerobic metabolism during eukaryotic evolution. *Sci Adv* **6**, eabb7258 (2020).

139. Y. Hoshino, E. A. Gaucher, Evolution of bacterial steroid biosynthesis and its impact on eukaryogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2101276118 (2021).

140. N. Cermakian, T. M. Ikeda, P. Miramontes, B. F. Lang, M. W. Gray, R. Cedergren, On the evolution of the single-subunit RNA polymerases. *J. Mol. Evol.* **45**, 671–681 (1997).

141. J. Filée, P. Forterre, Viral proteins functioning in organelles: a cryptic origin? *Trends Microbiol.* **13**, 510–513 (2005).

142. T. E. Shutt, M. W. Gray, Bacteriophage origins of mitochondrial replication and transcription proteins. *Trends Genet.* **22**, 90–95 (2006).

143. R. Harada, Y. Inagaki, Phage origin of mitochondrion-localized family A DNA polymerases in kinetoplastids and diplonemids. *Genome Biol. Evol.* **13** (2021).

144. A. J. Roger, E. Susko, M. M. Leger, Evolution: Reconstructing the timeline of eukaryogenesis. *Curr. Biol.* **31**, R193–R196 (2021).

145. F. Burki, M. M. Sandin, M. Jamy, Diversity and ecology of protists revealed by metabarcoding. *Curr. Biol.* **31**, R1267–R1280 (2021).

146. E. Alacid, T. A. Richards, A cell-cell atlas approach for understanding symbiotic interactions between microbes. *Curr. Opin. Microbiol.* **64**, 47–59 (2021).

147. R. M. R. Gawryluk, D. V. Tikhonenkov, E. Hehenberger, F. Husnik, A. P. Mylnikov, P. J. Keeling, Non-photosynthetic predators are sister to red algae. *Nature* **572**, 240–243 (2019).

148. M. E. Schön, V. V. Zlatogursky, R. P. Singh, C. Poirier, S. Wilken, V. Mathur, J. F. H. Strassert, J. Pinhassi, A. Z. Worden, P. J. Keeling, T. J. G. Ettema, J. G. Wideman, F. Burki, Single cell genomics reveals plastid-lacking Picozoa are close relatives of red algae. *Nat. Commun.* **12**, 6651 (2021).

149. C. W. Stairs, P. Táborský, E. D. Salomaki, M. Kolisko, T. Pánek, L. Eme, M. Hradilová, Č. Vlček, J. Jerlström-Hultqvist, A. J. Roger, I. Čepička, Anaeramoebae are a divergent lineage of eukaryotes that shed light on the transition from anaerobic mitochondria to hydrogenosomes. *Curr. Biol.* **31**, 5605-5612. e5 (2021).

150. M. W. Brown, A. A. Heiss, R. Kamikawa, Y. Inagaki, A. Yabuki, A. K. Tice, T. Shiratori, K.-I. Ishida, T. Hashimoto, A. G. B. Simpson, A. J. Roger, Phylogenomics places orphan protistan lineages in a novel eukaryotic super-group. *Genome Biol. Evol.* **10**, 427–433 (2018).

151. J. Janouškovec, D. V. Tikhonenkov, F. Burki, A. T. Howe, F. L. Rohwer, A. P. Mylnikov, P. J. Keeling, A new lineage of eukaryotes illuminates early mitochondrial genome reduction. *Curr. Biol.* **27**, 3717-3724.e5 (2017).

152. G. Lax, Y. Eglit, L. Eme, E. M. Bertrand, A. J. Roger, A. G. B. Simpson, Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. *Nature* **564**, 410–414 (2018).

153. G. Torruella, A. de Mendoza, X. Grau-Bové, M. Antó, M. A. Chaplin, J. del Campo, L. Eme, G. Pérez-Cordón, C. M. Whipps, K. M. Nichols, R. Paley, A. J. Roger, A. Sitjà-Bobadilla, S. Donachie, I. Ruiz-Trillo, Phylogenomics reveals convergent evolution of lifestyles in close relatives of animals and fungi. *Curr. Biol.* **25**, 2404–2410 (2015).

154. A. A. Heiss, M. Kolisko, F. Ekelund, M. W. Brown, A. J. Roger, A. G. B. Simpson, Combined morphological and phylogenomic re-examination of malawimonads, a critical taxon for inferring the evolutionary history of eukaryotes. *R. Soc. Open Sci.* **5**, 171707 (2018).

155. L. W. Parfrey, D. J. G. Lahr, A. H. Knoll, L. A. Katz, Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 13624–13629 (2011).

156. L. Eme, S. C. Sharpe, M. W. Brown, A. J. Roger, On the age of eukaryotes: evaluating evidence from fossils and molecular clocks. *Cold Spring Harb. Perspect. Biol.* **6** (2014).

157. H. C. Betts, M. N. Puttick, J. W. Clark, T. A. Williams, P. C. J. Donoghue, D. Pisani, Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nat Ecol Evol* **2**, 1556–1562 (2018).

158. V. L. Koumandou, B. Wickstead, M. L. Ginger, M. van der Giezen, J. B. Dacks, M. C. Field, Molecular paleontology and complexity in the last eukaryotic common ancestor. *Crit. Rev. Biochem. Mol. Biol.* **48**, 373–396 (2013).

159. E. S. Deutekom, B. Snel, T. J. P. van Dam, Benchmarking orthology methods using phylogenetic patterns defined at the base of Eukaryotes. *Brief. Bioinform.* **22** (2021).

160. D. He, O. Fiz-Palacios, C.-J. Fu, J. Fehling, C.-C. Tsai, S. L. Baldauf, An alternative root for the eukaryote tree of life. *Curr. Biol.* **24**, 465–470 (2014).

161. R. Derelle, G. Torruella, V. Klimeš, H. Brinkmann, E. Kim, Č. Vlček, B. F. Lang, M. Eliáš, Bacterial proteins pinpoint a single eukaryotic root. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E693-9 (2015).

162. L. A. Katz, J. R. Grant, L. W. Parfrey, J. G. Burleigh, Turning the crown upside down: gene tree parsimony roots the eukaryotic tree of life. *Syst. Biol.* **61**, 653–660 (2012).

163. M. A. Ceron Romero, M. M. Fonseca, L. de O. Martins, D. Posada, L. A. Katz, Phylogenomic analyses of 2,786 genes in 158 lineages support a root of the eukaryotic tree of life between opisthokonts (animals, Fungi and their microbial relatives) and all other lineages, *bioRxiv* (2021). https://doi.org/10.1101/2021.02.26.433005.

164. K. More, C. M. Klinger, L. D. Barlow, J. B. Dacks, Evolution and natural history of membrane trafficking in eukaryotes. *Curr. Biol.* **30**, R553–R564 (2020).

165. M. M. Leger, M. Petrů, V. Žárský, L. Eme, Č. Vlček, T. Harding, B. F. Lang, M. Eliáš, P. Doležal, A. J. Roger, An ancestral bacterial division system is widespread in eukaryotic mitochondria. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 10239–10246 (2015).

166. M. Müller, M. Mentel, J. J. van Hellemond, K. Henze, C. Woehle, S. B. Gould, R.-Y. Yu, M. van der Giezen, A. G. M. Tielens, W. F. Martin, Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. *Microbiol. Mol. Biol. Rev.* **76**, 444–495 (2012).

167. R. M. R. Gawryluk, C. W. Stairs, Diversity of electron transport chains in anaerobic protists. *Biochim. Biophys. Acta Bioenerg.* **1862**, 148334 (2021).

168. C. W. Stairs, M. M. Leger, A. J. Roger, Diversity and origins of anaerobic metabolism in mitochondria and related organelles. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140326 (2015).

169. A. Karnkowska, V. Vacek, Z. Zubáčová, S. C. Treitli, R. Petrželková, L. Eme, L. Novák, V. Žárský, L. D. Barlow, E. K. Herman, P. Soukal, M. Hroudová, P. Doležal, C. W. Stairs, A. J. Roger, M. Eliáš, J. B. Dacks, Č. Vlček, V. Hampl, A eukaryote without a mitochondrial organelle. *Curr. Biol.* **26**, 1274–1284 (2016).

170. L. A. Katz, Recent events dominate interdomain lateral gene transfers between prokaryotes and eukaryotes and, with the exception of endosymbiotic gene transfers, few ancient transfer events persist. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140324 (2015).

171. W. F. Martin, Too much eukaryote LGT. *Bioessays* **39**, 1700115 (2017).

172. M. M. Leger, L. Eme, C. W. Stairs, A. J. Roger, Demystifying eukaryote lateral gene transfer (response to Martin 2017 DOI: 10.1002/bies.201700115). *Bioessays* **40**, 1700242 (2018).

173. S. J. Sibbald, L. Eme, J. M. Archibald, A. J. Roger, Lateral Gene Transfer Mechanisms and Pan-genomes in Eukaryotes. *Trends Parasitol.* **36**, 927–941 (2020).

174. R. I. Ponce-Toledo, P. Deschamps, P. López-García, Y. Zivanovic, K. Benzerara, D. Moreira, An Early-Branching Freshwater Cyanobacterium at the Origin of Plastids. *Curr. Biol.* **27**, 386–391 (2017).

175. K. R. Moore, C. Magnabosco, L. Momper, D. A. Gold, T. Bosak, G. P. Fournier, An expanded ribosomal phylogeny of Cyanobacteria supports a deep placement of plastids. *Front. Microbiol.* **10**, 1612 (2019).

176. P. Sánchez-Baracaldo, J. A. Raven, D. Pisani, A. H. Knoll, Early photosynthetic eukaryotes inhabited low-salinity habitats. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E7737–E7745 (2017).

177. J. F. H. Strassert, I. Irisarri, T. A. Williams, F. Burki, A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. *Nat. Commun.* **12**, 1879 (2021).

178. E. C. M. Nowack, M. Melkonian, G. Glöckner, Chromatophore genome sequence of Paulinella sheds light on acquisition of photosynthesis by eukaryotes. *Curr. Biol.* **18**, 410–418 (2008).

179. T. Nakayama, K.-I. Ishida, Another acquisition of a primary photosynthetic organelle is underway in Paulinella chromatophora. *Curr. Biol.* **19**, R284-5 (2009).

180. B. Marin, E. C. M. Nowack, M. Melkonian, A plastid in the making: evidence for a second primary endosymbiosis. *Protist* **156**, 425–432 (2005).

181. L. Delaye, C. Valadez-Cano, B. Pérez-Zamorano, How really ancient is Paulinella chromatophora? *PLoS Curr.*, doi: 10.1371/currents.tol.e68a099364bb1a1e129a17b4e06b0c6b (2016).

182. S. J. Sibbald, J. M. Archibald, Genomic insights into Plastid evolution. *Genome Biol. Evol.* **12**, 978–990 (2020).

183. E. V. Koonin, M. Krupovic, V. I. Agol, The Baltimore Classification of Viruses 50 Years Later: How Does It Stand in the Light of Virus Evolution? *Microbiol. Mol. Biol. Rev.* **85**, e0005321 (2021).

184. D. Prangishvili, D. H. Bamford, P. Forterre, J. Iranzo, E. V. Koonin, M. Krupovic, The enigmatic archaeal virosphere. *Nat. Rev. Microbiol.* **15**, 724–739 (2017).

185. M. Krupovic, V. Cvirkaite-Krupovic, J. Iranzo, D. Prangishvili, E. V. Koonin, Viruses of archaea: Structural, functional, environmental and evolutionary genomics. *Virus Res.* **244**, 181–193 (2018).

186. P. López-García, The place of viruses in biology in light of the metabolism- versus-replication-first debate. *Hist. Philos. Life Sci.* **34**, 391–406 (2012).

187. E. V. Koonin, M. Krupovic, S. Ishino, Y. Ishino, The replication machinery of LUCA: common origin of DNA replication and transcription. *BMC Biol.* **18**, 61 (2020).

188. N. A. T. Irwin, A. A. Pittis, T. A. Richards, P. J. Keeling, Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. *Nat. Microbiol.* **7**, 327–336 (2022).

189. E. V. Koonin, M. Krupovic, The depths of virus exaptation. *Curr. Opin. Virol.* **31**, 1–8 (2018).

190. E. V. Koonin, V. V. Dolja, M. Krupovic, Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology* **479–480**, 2–25 (2015).

191. P. Forterre, To be or not to be alive: How recent discoveries challenge the traditional definitions of viruses and life. *Stud. Hist. Philos. Biol. Biomed. Sci.* **59**, 100–108 (2016).

192. J. Guglielmini, A. C. Woo, M. Krupovic, P. Forterre, M. Gaia, Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 19585–19592 (2019).

193. M. Moniruzzaman, A. R. Weinheimer, C. A. Martinez-Gutierrez, F. O. Aylward, Widespread endogenization of giant viruses shapes genomes of green algae. *Nature* **588**, 141–145 (2020).

194. A. B. Collens, L. A. Katz, Opinion: Genetic conflict with mobile elements drives eukaryotic genome evolution, and perhaps also eukaryogenesis. *J. Hered.* **112**, 140–144 (2021).

195. D. Raoult, S. Audic, C. Robert, C. Abergel, P. Renesto, H. Ogata, B. La Scola, M. Suzan, J.-M. Claverie, The 1.2-megabase genome sequence of Mimivirus. *Science* **306**, 1344–1350 (2004).

196. F. Schulz, N. Yutin, N. N. Ivanova, D. R. Ortega, T. K. Lee, J. Vierheilig, H. Daims, M. Horn, M. Wagner, G. J. Jensen, N. C. Kyrpides, E. V. Koonin, T. Woyke, Giant viruses with an expanded complement of translation system components. *Science* **356**, 82–85 (2017).

197. J. Abrahão, L. Silva, L. S. Silva, J. Y. B. Khalil, R. Rodrigues, T. Arantes, F. Assis, P. Boratto, M. Andrade, E. G. Kroon, B. Ribeiro, I. Bergier, H. Seligmann, E. Ghigo, P. Colson, A. Levasseur, G. Kroemer, D. Raoult, B. La Scola, Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. *Nat. Commun.* **9**, 749 (2018).

198. C. R. Schvarcz, G. F. Steward, A giant virus infecting green algae encodes key fermentation genes. *Virology* **518**, 423–433 (2018).

199. E. V. Koonin, N. Yutin, Evolution of the large nucleocytoplasmic DNA viruses of eukaryotes and convergent origins of viral gigantism. *Adv. Virus Res.* **103**, 167–202 (2019).

200. G. Yildirir, J. Sperschneider, M. Malar C, E. C. H. Chen, W. Iwasaki, C. Cornell, N. Corradi, Long reads and Hi-C sequencing illuminate the two-compartment genome of the model arbuscular mycorrhizal symbiont *Rhizophagus irregularis*. *New Phytol.* **233**, 1097–1107 (2022).

201. V. Da Cunha, M. Gaia, H. Ogata, O. Jaillon, T. O. Delmont, P. Forterre, Giant viruses encode actin-related proteins. *Mol. Biol. Evol.* **39** (2022).

202. S. Kijima, T. O. Delmont, U. Miyazaki, M. Gaia, H. Endo, H. Ogata, Discovery of viral myosin genes with complex evolutionary history within plankton. *Front. Microbiol.* **12**, 683294 (2021).

203. G. Yoshikawa, R. Blanc-Mathieu, C. Song, Y. Kayama, T. Mochizuki, K. Murata, H. Ogata, M. Takemura, Medusavirus, a novel large DNA virus discovered from hot spring water. *J. Virol.* **93** (2019).

204. R. R. Novoa, G. Calderita, R. Arranz, J. Fontana, H. Granzow, C. Risco, Virus factories: associations of cell organelles for viral replication and morphogenesis. *Biol. Cell* **97**, 147–172 (2005).

205. M. Suzan-Monti, B. La Scola, L. Barrassi, L. Espinosa, D. Raoult, Ultrastructural characterization of the giant volcano-like virus factory of Acanthamoeba polyphaga Mimivirus. *PLoS One* **2**, e328 (2007).

206. M. Krupovic, J. H. Kuhn, M. G. Fischer, A classification system for virophages and satellite viruses. *Arch. Virol.* **161**, 233–247 (2016).

207. T. A. Williams, T. M. Embley, E. Heinz, Informational gene phylogenies do not support a fourth domain of life for nucleocytoplasmic large DNA viruses. *PLoS One* **6**, e21080 (2011).

208. D. Moreira, P. López-García, Evolution of viruses and cells: do we need a fourth domain of life to explain the origin of eukaryotes? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140327 (2015).

209. E. V. Koonin, N. Yutin, Multiple evolutionary origins of giant viruses. *F1000Res.* **7**, 1840 (2018).

210. D. Bäckström, N. Yutin, S. L. Jørgensen, J. Dharamshi, F. Homa, K. Zaremba-Niedwiedzka, A. Spang, Y. I. Wolf, E. V. Koonin, T. J. G. Ettema, Virus genomes from deep sea sediments expand the ocean megavirome and support independent origins of viral gigantism. *MBio* **10** (2019).

211. P. Forterre, M. Krupovic, D. Prangishvili, Cellular domains and viral lineages. *Trends Microbiol.* **22**, 554–558 (2014).

212. M. Takemura, Medusavirus ancestor in a proto-eukaryotic cell: Updating the hypothesis for the viral origin of the nucleus. *Front. Microbiol.* **11**, 571831 (2020).

213. P. López-García, L. Eme, D. Moreira, Symbiosis in eukaryotic evolution. (2017). https://doi.org/10.1016/j.jtbi.2017.02.031.

214. V. Chaikeeratisak, K. Nguyen, K. Khanna, A. F. Brilot, M. L. Erb, J. K. C. Coker, A. Vavilina, G. L. Newton, R. Buschauer, K. Pogliano, E. Villa, D. A. Agard, J. Pogliano, Assembly of a nucleus-like structure during viral replication in bacteria. *Science* **355**, 194–197 (2017).

215. C. Feschotte, C. Gilbert, Endogenous viruses: insights into viral evolution and impact on host biology. *Nat. Rev. Genet.* **13**, 283–296 (2012).

216. S. Medvedeva, J. Sun, N. Yutin, E. V. Koonin, T. Nunoura, C. Rinke, M. Krupovic, Viruses of Asgard archaea, *bioRxiv* (2021). https://doi.org/10.1101/2021.07.29.453957.

217. I. M. Rambo, V. de Anda, M. V. Langwig, B. J. Baker, Unique viruses that infect Archaea related to eukaryotes, *bioRxiv* (2021). https://doi.org/10.1101/2021.07.29.454249.

218. D. Tamarit, E. F. Caceres, M. Krupovic, R. Nijland, L. Eme, N. P. Robinson, T. J. G. Ettema, A closed Candidatus Odinarchaeum genome exposes Asgard archaeal viruses, *bioRxiv* (2021). https://doi.org/10.1101/2021.09.01.458545.

219. D. A. Liberles, B. Chang, K. Geiler-Samerotte, A. Goldman, J. Hey, B. Kaçar, M. Meyer, W. Murphy, D. Posada, A. Storfer, Emerging frontiers in the study of Molecular Evolution. *J. Mol. Evol.* **88**, 211–226 (2020).

220. A. Som, Causes, consequences and solutions of phylogenetic incongruence. *Brief. Bioinform.* **16**, 536–548 (2015).

221. E. A. Eloe-Fadrosh, N. N. Ivanova, T. Woyke, N. C. Kyrpides, Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nat Microbiol* **1**, 15032 (2016).

222. E. A. Eloe-Fadrosh, D. Paez-Espino, J. Jarett, P. F. Dunfield, B. P. Hedlund, A. E. Dekas, S. E. Grasby, A. L. Brady, H. Dong, B. R. Briggs, W.-J. Li, D. Goudeau, R. Malmstrom, A. Pati, J. Pett-Ridge, E. M. Rubin, T. Woyke, N. C. Kyrpides, N. N. Ivanova, Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. *Nat. Commun.* **7**, 10476 (2016).

223. N. C. Kyrpides, E. A. Eloe-Fadrosh, N. N. Ivanova, Microbiome Data Science: Understanding Our Microbial Planet. *Trends Microbiol.* **24**, 425–427 (2016).

224. D. H. Parks, C. Rinke, M. Chuvochina, P.-A. Chaumeil, B. J. Woodcroft, P. N. Evans, P. Hugenholtz, G. W. Tyson, Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* **2**, 1533–1542 (2017).

225. X. Liao, M. Li, Y. Zou, F.-X. Wu, Yi-Pan, J. Wang, Current challenges and solutions of *de novo* assembly. *Quant. Biol.* **7**, 90–109 (2019).

226. Y. Wang, Y. Zhao, A. Bollas, Y. Wang, K. F. Au, Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **39**, 1348–1365 (2021).

227. D. Bertrand, J. Shaw, M. Kalathiyappan, A. H. Q. Ng, M. S. Kumar, C. Li, M. Dvornicic, J. P. Soldo, J. Y. Koh, C. Tong, O. T. Ng, T. Barkham, B. Young, K. Marimuthu, K. R. Chng, M. Sikic, N. Nagarajan, Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat. Biotechnol.* **37**, 937–944 (2019).

228. M. Kolmogorov, D. M. Bickhart, B. Behsaz, A. Gurevich, M. Rayko, S. B. Shin, K. Kuhn, J. Yuan, E. Polevikov, T. P. L. Smith, P. A. Pevzner, metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).

229. L.-X. Chen, K. Anantharaman, A. Shaiber, A. M. Eren, J. F. Banfield, Accurate and complete genomes from metagenomes. *Genome Res.* **30**, 315–333 (2020).

230. M. R. Olm, A. Crits-Christoph, K. Bouma-Gregson, B. A. Firek, M. J. Morowitz, J. F. Banfield, inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol.* **39**, 727–736 (2021).

231. C. Quince, S. Nurk, S. Raguideau, R. James, O. S. Soyer, J. K. Summers, A. Limasset, A. M. Eren, R. Chikhi, A. E. Darling, STRONG: metagenomics strain resolution on assembly graphs. *Genome Biol.* **22**, 214 (2021).

232. B. Al-Shayeb, R. Sachdeva, L.-X. Chen, F. Ward, P. Munk, A. Devoto, C. J. Castelle, M. R. Olm, K. Bouma-Gregson, Y. Amano, C. He, R. Méheust, B. Brooks, A. Thomas, A. Lavy, P. Matheus-Carnevali, C. Sun, D. S. A. Goltsman, M. A. Borton, A. Sharrar, A. L. Jaffe, T. C. Nelson, R. Kantor, R. Keren, K. R. Lane, I. F. Farag, S. Lei, K. Finstad, R. Amundson, K. Anantharaman, J. Zhou, A. J. Probst, M. E. Power, S. G. Tringe, W.-J. Li, K. Wrighton, S. Harrison, M. Morowitz, D. A. Relman, J. A. Doudna, A.-C. Lehours, L. Warren, J. H. D. Cate, J. M. Santini, J. F. Banfield, Clades of huge phages from across Earth's ecosystems. *Nature* **578**, 425–431 (2020).

233. S. J. Sibbald, J. M. Archibald, More protist genomes needed. *Nat. Ecol. Evol.* **1**, 145 (2017).

234. C. L. McGrath, L. A. Katz, Genome diversity in microbial eukaryotes. *Trends Ecol. Evol.* **19**, 32–38 (2004).

235. P. T. West, A. J. Probst, I. V. Grigoriev, B. C. Thomas, J. F. Banfield, Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res.* **28**, 569–580 (2018).

236. A. Karnkowska, S. C. Treitli, O. Brzoň, L. Novák, V. Vacek, P. Soukal, L. D. Barlow, E. K. Herman, S. V. Pipaliya, T. Pánek, D. Žihala, R. Petrželková, A. Butenko, L. Eme, C. W. Stairs, A. J. Roger, M. Eliáš, J. B. Dacks, V. Hampl, The oxymonad genome displays canonical eukaryotic complexity in the absence of a mitochondrion. *Mol. Biol. Evol.* **36**, 2292–2312 (2019).

237. T. A. Williams, D. Schrempf, G. J. Szöllősi, C. J. Cox, P. G. Foster, T. M. Embley, Inferring the Deep Past from Molecular Data. *Genome Biol. Evol.* **13** (2021).

238. S. M. Crotty, B. Q. Minh, N. G. Bean, B. R. Holland, J. Tuke, L. S. Jermiin, A. V. Haeseler, GHOST: Recovering historical signal from heterotachously evolved sequence alignments. *Syst. Biol.* **69**, 249–264 (2020).

239. L. Salichos, A. Rokas, Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497**, 327–331 (2013).

240. R. C. Thomson, J. M. Brown, On the need for new measures of phylogenomic support. *Syst. Biol.* **71**, 917–920 (2022).

241. K. Kobert, L. Salichos, A. Rokas, A. Stamatakis, Computing the internode certainty and related measures from partial gene trees. *Mol. Biol. Evol.* **33**, 1606–1617 (2016).

242. M. Long, N. W. VanKuren, S. Chen, M. D. Vibranovski, New gene evolution: little did we know. *Annu. Rev. Genet.* **47**, 307–333 (2013).

243. D. I. Andersson, J. Jerlström-Hultqvist, J. Näsvall, Evolution of new functions de novo and from preexisting genes. *Cold Spring Harb. Perspect. Biol.* **7**, a017996 (2015).

244. T. Dagan, Phylogenomic networks. *Trends Microbiol.* **19**, 483–491 (2011).

245. J. M. Chan, G. Carlsson, R. Rabadan, Topology of viral evolution. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 18566–18571 (2013).

246. P. G. Cámara, Topological methods for genomics: present and future directions. *Curr. Opin. Syst. Biol.* **1**, 95–101 (2017).

247. B. Morel, P. Schade, S. Lutteropp, T. A. Williams, G. J. Szöllősi, A. Stamatakis, SpeciesRax: A Tool for Maximum Likelihood Species Tree Inference from Gene Family Trees under Duplication, Transfer, and Loss. *Mol. Biol. Evol.* **39**, msab365 (2022).

248. A. A. Davín, E. Tannier, T. A. Williams, B. Boussau, V. Daubin, G. J. Szöllősi, Gene transfers can date the tree of life. *Nat Ecol Evol* **2**, 904–909 (2018).

249. J. M. Wolfe, G. P. Fournier, Horizontal gene transfer constrains the timing of methanogen evolution. *Nat Ecol Evol* **2**, 897–903 (2018).

250. L. Orlando, M. T. P. Gilbert, E. Willerslev, Reconstructing ancient genomes and epigenomes. *Nat. Rev. Genet.* **16**, 395–408 (2015).

251. M. Leonardi, P. Librado, C. Der Sarkissian, M. Schubert, A. H. Alfarhan, S. A. Alquraishi, K. A. S. Al-Rasheid, C. Gamba, E. Willerslev, L. Orlando, Evolutionary patterns and processes: Lessons from ancient DNA. *Syst. Biol.*, syw059 (2016).

252. E. Cappellini, A. Prohaska, F. Racimo, F. Welker, M. W. Pedersen, M. E. Allentoft, P. de Barros Damgaard, P. Gutenbrunner, J. Dunne, S. Hammann, M. Roffet-Salque, M. Ilardo, J. V. Moreno-Mayar, Y. Wang, M. Sikora, L. Vinner, J. Cox, R. P. Evershed, E. Willerslev, Ancient biomolecules and evolutionary inference. *Annu. Rev. Biochem.* **87**, 1029–1060 (2018).

253. C. Pont, S. Wagner, A. Kremer, L. Orlando, C. Plomion, J. Salse, Paleogenomics: reconstruction of plant evolutionary trajectories from modern and ancient DNA. *Genome Biol.* **20** (2019).

254. L. A. Arriola, A. Cooper, L. S. Weyrich, Palaeomicrobiology: Application of ancient DNA sequencing to better understand bacterial genome evolution and adaptation. *Front. Ecol. Evol.* **8** (2020).

255. Y. Lammers, P. D. Heintzman, I. G. Alsos, Environmental palaeogenomic reconstruction of an Ice Age algal population. *Commun. Biol.* **4**, 220 (2021).

256. R. Liang, Z. Li, M. C. Y. Lau Vetter, T. A. Vishnivetskaya, O. G. Zanina, K. G. Lloyd, S. M. Pfiffner, E. M. Rivkina, W. Wang, J. Wiggins, J. Miller, R. L. Hettich, T. C. Onstott, Genomic reconstruction of fossil and living microorganisms in ancient Siberian permafrost. *Microbiome* **9**, 110 (2021).

257. K. E. Omland, The assumptions and challenges of ancestral state reconstructions. *Syst. Biol.* **48**, 604–611 (1999).

258. J. B. Joy, R. H. Liang, R. M. McCloskey, T. Nguyen, A. F. Y. Poon, Ancestral reconstruction. *PLoS Comput. Biol.* **12**, e1004763 (2016).

259. R. Merkl, R. Sterner, Ancestral protein reconstruction: techniques and applications. *Biol. Chem.* **397**, 1–21 (2016).

260. Y. Gumulya, E. M. J. Gillam, Exploring the past and the future of protein evolution with ancestral sequence reconstruction: the 'retro' approach to protein engineering. *Biochem. J.* **474**, 1–19 (2017).

261. A. G. A. Selberg, E. A. Gaucher, D. A. Liberles, Ancestral sequence reconstruction: From chemical paleogenetics to maximum likelihood algorithms and beyond. *J. Mol. Evol.* **89**, 157–164 (2021).

262. J. W. Thornton, Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat. Rev. Genet.* **5**, 366–375 (2004).

263. G. K. A. Hochberg, J. W. Thornton, Reconstructing ancient proteins to understand the causes of structure and function. *Annu. Rev. Biophys.* **46**, 247–269 (2017).

264. M. L. Mascotti, Resurrecting enzymes by ancestral Sequence Reconstruction. *Methods Mol. Biol.* **2397**, 111–136 (2022).

265. G. C. Finnigan, V. Hanson-Smith, T. H. Stevens, J. W. Thornton, Evolution of increased complexity in a molecular machine. *Nature* **481**, 360–364 (2012).

266. P. M. Shih, A. Occhialini, J. C. Cameron, P. J. Andralojc, M. A. J. Parry, C. A. Kerfeld, Biochemical characterization of predicted Precambrian RuBisCO. *Nat. Commun.* **7**, 10382 (2016).

267. M. A. Siddiq, G. K. Hochberg, J. W. Thornton, Evolution of protein specificity: insights from ancestral protein reconstruction. *Curr. Opin. Struct. Biol.* **47**, 113–122 (2017).

268. A. S. Pillai, S. A. Chandler, Y. Liu, A. V. Signore, C. R. Cortez-Romero, J. L. P. Benesch, A. Laganowsky, J. F. Storz, G. K. A. Hochberg, J. W. Thornton, Origin of complexity in haemoglobin evolution. *Nature* **581**, 480–485 (2020).

269. P. Schwille, J. Spatz, K. Landfester, E. Bodenschatz, S. Herminghaus, V. Sourjik, T. J. Erb, P. Bastiaens, R. Lipowsky, A. Hyman, P. Dabrock, J.-C. Baret, T. Vidakovic-Koch, P. Bieling, R. Dimova, H. Mutschler, T. Robinson, T.-Y. D. Tang, S. Wegner, K. Sundmacher, MaxSynBio: Avenues towards creating cells from the bottom up. *Angew. Chem. Int. Ed Engl.* **57**, 13382–13392 (2018).

270. I. V. Surovtsev, C. Jacobs-Wagner, Subcellular organization: A critical feature of bacterial cell replication. *Cell* **172**, 1271–1293 (2018).

271. J. L. S. Milne, M. J. Borgnia, A. Bartesaghi, E. E. H. Tran, L. A. Earl, D. M. Schauder, J. Lengyel, J. Pierson, A. Patwardhan, S. Subramaniam, Cryo-electron microscopy – a primer for the non-microscopist. *FEBS J.* **280**, 28–45 (2013).

272. M. Beck, W. Baumeister, Cryo-electron tomography: Can it reveal the molecular sociology of cells in atomic detail? *Trends Cell Biol.* **26**, 825–837 (2016).

273. C. M. Oikonomou, G. J. Jensen, Cellular electron cryotomography: Toward structural biology in situ. *Annu. Rev. Biochem.* **86**, 873–896 (2017).

274. M. J. Dobro, C. M. Oikonomou, A. Piper, J. Cohen, K. Guo, T. Jensen, J. Tadayon, J. Donermeyer, Y. Park, B. A. Solis, A. Kjær, A. I. Jewett, A. W. McDowall, S. Chen, Y.-W. Chang, J. Shi, P. Subramanian, C. V. Iancu, Z. Li, A. Briegel, E. I. Tocheva, M. Pilhofer, G. J. Jensen, Uncharacterized bacterial structures revealed by electron cryotomography. *J. Bacteriol.* **199** (2017).

275. C. Greening, T. Lithgow, Formation and function of bacterial organelles. *Nat. Rev. Microbiol.* **18**, 677–689 (2020).

276. C. Seeger, K. Dyrhage, M. Mahajan, A. Odelgard, S. B. Lind, S. G. E. Andersson, The subcellular proteome of a Planctomycetes bacterium shows that newly evolved proteins have distinct fractionation patterns. *Front. Microbiol.* **12**, 643045 (2021).

277. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

278. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

279. I. R. Humphreys, J. Pei, M. Baek, A. Krishnakumar, I. Anishchenko, S. Ovchinnikov, J. Zhang, T. J. Ness, S. Banjade, S. R. Bagde, V. G. Stancheva, X.-H. Li, K. Liu, Z. Zheng, D. J. Barrero, U. Roy, J. Kuper, I. S. Fernández, B. Szakal, D. Branzei, J. Rizo, C. Kisker, E. C. Greene, S. Biggins, S. Keeney, E. A. Miller, J. C. Fromme, T. L. Hendrickson, Q. Cong, D. Baker, Computed structures of core eukaryotic protein complexes. *Science* **374**, eabm4805 (2021).

280. A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, J. Moult, Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins* **89**, 1607–1617 (2021).

281. M. Gupta, C. M. Azumaya, M. Moritz, S. Pourmal, A. Diallo, G. E. Merz, G. Jang, M. Bouhaddou, A. Fossati, A. F. Brilot, D. Diwanji, E. Hernandez, N. Herrera, H. T. Kratochvil, V. L. Lam, F. Li, Y. Li, H. C. Nguyen, C. Nowotny, T. W. Owens, J. K. Peters, A. N. Rizo, U. Schulze-Gahmen, A. M. Smith, I. D. Young, Z. Yu, D. Asarnow, C. Billesbølle, M. G. Campbell, J. Chen, K.-H. Chen, U. S. Chio, M. S. Dickinson, L. Doan, M. Jin, K. Kim, J. Li, Y.-L. Li, E. Linossi, Y. Liu, M. Lo, J. Lopez, K. E. Lopez, A. Mancino, F. R. Moss 3rd, M. D. Paul, K. I. Pawar, A. Pelin, T. H. Pospiech Jr, C. Puchades, S. G. Remesh, M. Safari, K. Schaefer, M. Sun, M. C. Tabios, A. C. Thwin, E. W. Titus, R. Trenker, E. Tse, T. K. M. Tsui, F. Wang, K. Zhang, Y. Zhang, J. Zhao, F. Zhou, Y. Zhou, L. Zuliani-Alvarez, QCRG Structural Biology Consortium, D. A. Agard, Y. Cheng, J. S. Fraser, N. Jura, T. Kortemme, A. Manglik, D. R. Southworth, R. M. Stroud, D. L. Swaney, N. J. Krogan, A. Frost, O. S. Rosenberg, K. A. Verba, CryoEM and AI reveal a structure of SARS-CoV-2 Nsp2, a multifunctional protein involved in key host processes. *bioRxivorg*, doi: 10.1101/2021.05.10.443524 (2021).

282. G. Masrati, M. Landau, N. Ben-Tal, A. Lupas, M. Kosloff, J. Kosinski, Integrative structural biology in the era of accurate structure prediction. *J. Mol. Biol.* **433**, 167127 (2021).

283. T. M. Mayhew, Mapping the distributions and quantifying the labelling intensities of cell compartments by immunoelectron microscopy: progress towards a coherent set of methods. *J. Anat.* **219**, 647–660 (2011).

284. H. H. Tuson, J. S. Biteen, Unveiling the inner workings of live bacteria using super-resolution microscopy. *Anal. Chem.* **87**, 42–63 (2015).

285. L. Möckl, W. E. Moerner, Super-resolution microscopy with single molecules in biology and beyond-essentials, current trends, and future challenges. *J. Am. Chem. Soc.* **142**, 17828–17844 (2020).

286. S. Holden, Probing the mechanistic principles of bacterial cell division with super-resolution microscopy. *Curr. Opin. Microbiol.* **43**, 84–91 (2018).

287. N. Pende, A. Sogues, D. Megrian, A. Sartori-Rupp, P. England, H. Palabikyan, S. K.-M. R. Rittmann, M. Graña, A. M. Wehenkel, P. M. Alzari, S. Gribaldo, SepF is the FtsZ anchor in archaea, with features of an ancestral cell division system. *Nat. Commun.* **12**, 3214 (2021).

288. M. Kirschner, J. Gerhart, Evolvability. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 8420–8427 (1998).

289. M. Pigliucci, Is evolvability evolvable? *Nat. Rev. Genet.* **9**, 75–82 (2008).

290. J. L. Payne, A. Wagner, The causes of evolvability and their evolution. *Nat. Rev. Genet.* **20**, 24–38 (2019).

291. R. M. Alexander, The ideal and the feasible: physical constraints on evolution. *Biol. J. Linn. Soc. Lond.* **26**, 345–358 (1985).

292. J. M. Smith, R. Burian, S. Kauffman, P. Alberch, J. Campbell, B. Goodwin, R. Lande, D. Raup, L. Wolpert, Developmental constraints and evolution: A perspective from the Mountain Lake conference on development and evolution. *Q. Rev. Biol.* **60**, 265–287 (1985).

293. S. J. Arnold, Constraints on phenotypic evolution. *Am. Nat.* **140**, S85–S107 (1992).

294. C. Furusawa, N. Irie, Toward understanding of evolutionary constraints: experimental and theoretical approaches. *Biophys. Rev.* **12**, 1155–1161 (2020).

295. A. A. Sharov, Evolutionary constraints or opportunities? *Biosystems.* **123**, 9–18 (2014).

296. T. Garland Jr, Trade-offs. *Curr. Biol.* **24**, R60–R61 (2014).

297. L. Acerenza, Constraints, trade-offs and the currency of fitness. *J. Mol. Evol.* **82**, 117–127 (2016).

298. G. B. West, J. H. Brown, B. J. Enquist, A general model for the origin of allometric scaling laws in biology. *Science* **276**, 122–126 (1997).

299. G. B. West, W. H. Woodruff, J. H. Brown, Allometric scaling of metabolic rate from molecules and mitochondria to cells and mammals. *Proc. Natl. Acad. Sci. U. S. A.* **99 Suppl 1**, 2473–2478 (2002).

300. A. Giometto, F. Altermatt, F. Carrara, A. Maritan, A. Rinaldo, Scaling body size fluctuations. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 4646–4650 (2013).

301. J. A. G. M. de Visser, J. Hermisson, G. P. Wagner, L. Ancel Meyers, H. Bagheri-Chaichian, J. L. Blanchard, L. Chao, J. M. Cheverud, S. F. Elena, W. Fontana, G. Gibson, T. F. Hansen, D. Krakauer, R. C. Lewontin, C. Ofria, S. H. Rice, G. von Dassow, A. Wagner, M. C. Whitlock, Perspective: Evolution and detection of genetic robustness. *Evolution* **57**, 1959–1972 (2003).

302. H. Kitano, Towards a theory of biological robustness. *Mol. Syst. Biol.* **3**, 137 (2007).

303. J. Masel, M. V. Trotter, Robustness and evolvability. *Trends Genet.* **26**, 406–414 (2010).

304. M. Pigliucci, Genotype-phenotype mapping and the end of the "genes as blueprint" metaphor. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**, 557–566 (2010).

305. G. P. Wagner, J. Zhang, The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nat. Rev. Genet.* **12**, 204–213 (2011).

306. S. E. Ahnert, Structural properties of genotype–phenotype maps. *J. R. Soc. Interface* **14**, 20170275 (2017).

307. S. M. Prakadan, A. K. Shalek, D. A. Weitz, Scaling by shrinking: empowering single-cell "omics" with microfluidic devices. *Nat. Rev. Genet.* **18**, 345–361 (2017).

308. M. Adli, The CRISPR tool kit for genome editing and beyond. *Nat. Commun.* **9**, 1911 (2018).

309. J. Ohan, B. Pelle, P. Nath, J.-H. Huang, B. Hovde, M. Vuyisich, A. E. Dichosa, S. R. Starkenburg, High-throughput phenotyping of cell-to-cell interactions in gel microdroplet pico-cultures. *Biotechniques* **66**, 218–224 (2019).

310. T. Zahir, R. Camacho, R. Vitale, C. Ruckebusch, J. Hofkens, M. Fauvart, J. Michiels, High-throughput time-resolved morphology screening in bacteria reveals phenotypic responses to antibiotics. *Commun. Biol.* **2**, 269 (2019).

311. M. Acin-Albiac, P. Filannino, M. Gobbetti, R. Di Cagno, Microbial high throughput phenomics: The potential of an irreplaceable omics. *Comput. Struct. Biotechnol. J.* **18**, 2290–2299 (2020).

312. A.-K. Kaster, M. S. Sobol, Microbial single-cell omics: the crux of the matter. *Appl. Microbiol. Biotechnol.* **104**, 8209–8220 (2020).

313. N. S. McCarty, A. E. Graham, L. Studená, R. Ledesma-Amaro, Multiplexed CRISPR technologies for gene editing and transcriptional regulation. *Nat. Commun.* **11**, 1281 (2020).

314. R. D. Arroyo-Olarte, R. Bravo Rodríguez, E. Morales-Ríos, Genome editing in bacteria: CRISPR-Cas and beyond. *Microorganisms* **9**, 844 (2021).

315. B. E. Rubin, S. Diamond, B. F. Cress, A. Crits-Christoph, Y. C. Lou, A. L. Borges, H. Shivram, C. He, M. Xu, Z. Zhou, S. J. Smith, R. Rovinsky, D. C. J. Smock, K. Tang, T. K. Owens, N. Krishnappa, R. Sachdeva, R. Barrangou, A. M. Deutschbauer, J. F. Banfield, J. A. Doudna, Species- and site-specific genome editing in complex bacterial communities. *Nat. Microbiol.* **7**, 34–47 (2022).

316. S. G. Peisajovich, Evolutionary synthetic biology. *ACS Synth. Biol.* **1**, 199–210 (2012).

317. F. Baier, Y. Schaerli, "Addressing evolutionary questions with synthetic biology" in *Evolutionary Systems Biology* (Springer International Publishing, Cham, 2021), pp. 135–157.

318. T. Ijäs, R. Koskinen, Exploring biological possibility through synthetic biology. *Eur. J. Philos. Sci.* **11** (2021).

319. B. Van den Bergh, T. Swings, M. Fauvart, J. Michiels, Experimental design, population dynamics, and diversity in microbial experimental evolution. *Microbiol. Mol. Biol. Rev.* **82** (2018).

320. W. H. Lewis, G. Tahon, P. Geesink, D. Z. Sousa, T. J. G. Ettema, Innovations to culturing the uncultured microbial majority. *Nat. Rev. Microbiol.* **19**, 225–240 (2021).

321. S. M. Adl, D. Bass, C. E. Lane, J. Lukeš, C. L. Schoch, A. Smirnov, S. Agatha, C. Berney, M. W. Brown, F. Burki, P. Cárdenas, I. Čepička, L. Chistyakova, J. Del Campo, M. Dunthorn, B. Edvardsen, Y. Eglit, L. Guillou, V. Hampl, A. A. Heiss, M. Hoppenrath, T. Y. James, A. Karnkowska, S. Karpov, E. Kim, M. Kolisko, A. Kudryavtsev, D. J. G. Lahr, E. Lara, L. Le Gall, D. H. Lynn, D. G. Mann, R. Massana, E. A. D. Mitchell, C. Morrow, J. S. Park, J. W. Pawlowski, M. J. Powell, D. J. Richter, S. Rueckert, L. Shadwick, S. Shimano, F. W. Spiegel, G. Torruella, N. Youssef, V. Zlatogursky, Q. Zhang, Revisions to the classification, nomenclature, and diversity of eukaryotes. *J. Eukaryot. Microbiol.* **66**, 4–119 (2019).

322. J. F. H. Strassert, M. Jamy, A. P. Mylnikov, D. V. Tikhonenkov, F. Burki, New phylogenomic analysis of the enigmatic phylum Telonemia further resolves the eukaryote tree of life. *Mol. Biol. Evol.* **36**, 757–765 (2019).

323. E. Yazaki, A. Yabuki, A. Imaizumi, K. Kume, T. Hashimoto, Y. Inagaki, Phylogenomics invokes the clade housing Cryptista, Archaeplastida, and Microheliella maris, *bioRxiv* (2021). https://doi.org/10.1101/2021.08.29.458128.

CHAPTER 6

Archaea

Nina Dombrowski, Tara Mahendrarajah, Sarah T. Gross, Laura Eme, and Anja Spang

Practical Handbook of Microbiology, Fourth Edition, 2021

SUMMARY AND CONTRIBUTIONS

The primary goal of this book chapter was to provide a general overview of recently updated research on the biodiversity, taxonomy, evolution, and ecology of Archaea. The Archaea were formally proposed as a domain of life almost 50 years ago, with evidence that they are likely more closely related to eukaryotes despite sharing many superficial features with Bacteria. At the time this fundamentally altered the view of the tree of life (TOL) and became the foundation for retracing the evolutionary history of the prokaryotes and eukaryotes. In the years that followed, the expansion of the archaeal tree paralleled advances in environmental sampling technologies, gene and genome sequencing approaches, and complex phylogenetic inference. Initially viewed as extremophiles occupying narrow niches, the Archaea have since been found to be ubiquitous in the natural environment, making them key ecological players and drivers of biogeochemical cycling. Research into the taxonomy and phylogeny of the newly proposed archaeal lineages led to the identification of four major supergroups, the Euryarchaeota, TACK (Thaum-, Aig-, Cren-, and Korarchaeota), Asgard, and DPANN (Diapherotrites, Parv-, Aenigma-, Nano-, and Nanohaloarchaeota). Here, we briefly summarize key features of important members and newly discovered lineages belonging to these four groups. We discussed how the discovery of Asgard archaea has led to a better understanding of the origin of eukaryotes. Furthermore, we detail how updated phylogenetic analyses including the increased sampling of the biosphere also uncovered a large radiation of archaea with small cell and genome sizes, referred to as the DPANN. Aside from their ultrasmall morphological features, they have been shown to lack key biosynthetic pathways and the superphylum is believed to consist primarily of ectosymbionts that rely on a host to complement lacking metabolic requirements. The DPANN are often resolved as an early-branching clade within the archaeal domain, which has huge implications for early evolution and diversification of the Archaea. However, this position is still an open and unanswered question and must be addressed using a greater representation of archaeal genomes. Considering these new lineages, we close with an overview of the role of Archaea within the human microbiome, including the prevalence of methanogens identified in the oral cavity and gut to emerging evidence of archaeal colonization of other human habitats including the skin and lungs, among others. All authors were involved in writing and revision of the book chapter. My primary contributions included researching, analyzing, and writing the entire section on the “Human Archaeome”. I also worked with Nina Dombrowski and Anja Spang to generate Figs. 1 and 2.

CONTENTS

INTRODUCTION

Just about half a century ago, all prokaryotes, i.e., cells without nucleus, were classified within one kingdom: *Monera*. However, in the late 1970s, scientists were starting to recognize that this classification system, based predominantly on morphological and metabolic traits, underestimated the vast diversity of prokaryotic life. Around the same time, the pioneering work of Carl Woese and George Fox led to the discovery that prokaryotes were, in fact, composed of two fundamentally different domains of life—the *Bacteria* and the *Archaea* (originally referred to as "Eubacteria" and "Archaebacteria," respectively) (*1*). Woese and coworkers used the RNA components of the ribosome to reconstruct the first phylogenetic tree of life based on molecular data (*2*), which divided cellular organisms into three separate domains of life (Fig. 1A)—the *Bacteria*, *Archaea*, and *Eukarya*, the latter of which comprised all organisms with a true nucleus (*2*). At that time, it was suggested that *Archaea*, in spite of their superficial similarity to *Bacteria*, may be more closely related to eukaryotes than *Bacteria*. In fact, they seemed to harbor simplified versions of eukaryotic informational processing machineries (replication, transcription, translation, and cell division), in addition to unique characteristics such as ether-bound isoprenoids rather than ester-bound fatty acid-based lipids (Table 1). Subsequent research on *Archaea*, accompanied by extensive methodological developments in environmental microbiology, sequencing technologies, physiology, cell biology, and phylogenetics, has further changed our view on the diversity of life, the tree topology, as well as the ecological and evolutionary importance of *Archaea*. In particular, the use of cultivation-independent techniques, such as metagenomics and single-cell genomics, which allow us to obtain genomes of uncultivated organisms directly from environmental samples (*3*, *4*), have been a key element leading to our changed perception of archaeal diversity and distribution. While *Archaea* have originally been viewed as comprising predominantly "extremophilic" organisms inhabiting environments with high temperature, salinity, and high or low pH, they are now known to be ubiquitous in all environments on Earth, including marine waters and freshwater lakes, sediments, soils (including plant roots), aquifers, and the human microbiome to name a few (*5*–*7*). With their widespread ecological distribution and important metabolic capabilities, *Archaea* are recognized as key players in a wide variety of biogeochemical processes, including the sulfur, nitrogen, and carbon cycles (*8*). For instance, *Archaea* include the only known organisms able to conserve energy through the anaerobic production or consumption of methane in processes referred to as methanogenesis and anaerobic methane oxidation, respectively. Since methane is an extremely potent greenhouse gas, with a global-warming potential about 25 times greater than carbon dioxide, these *Archaea* have an essential role in the global carbon budget and consequently climate change (*9*). Finally, the study of archaeal phylogenetic diversity and evolution has fundamentally changed our understanding of the eukaryotic cell (see below) (*10*).

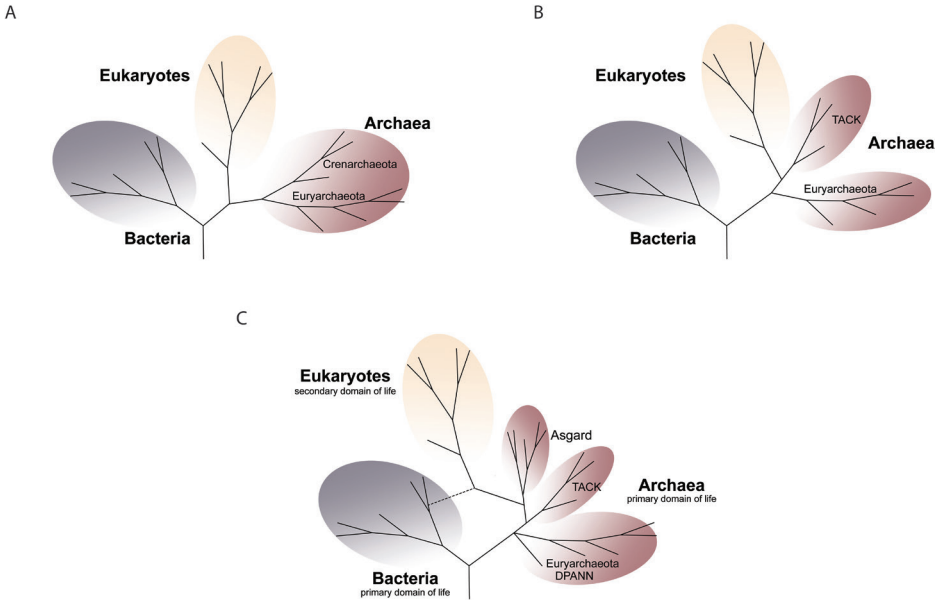


Fig. 1 | Schematic depictions of the relationship of *Archaea* with *Bacteria* and eukaryotes in the tree of life. A: Upon the discovery of *Archaea* as a separate domain, the tree of life was divided into three major domains. B: However, phylogenetic analyses of core informational proteins suggested later that eukaryotes may have evolved from within the *Archaea*, challenging the three-domain topology. C: Recent research, among others enabled by the discovery of the Asgard archaea, has shed further support on the branching of eukaryotes from within the *Archaea* (in terms of universal marker proteins). In turn, it has been suggested that the tree of life has two primary domains of life—the *Archaea* and *Bacteria*—and one secondary domain of life, which evolved from the former (see text for more details).

ARCHAEA AND THE TREE OF LIFE

Since the discovery of the *Archaea* as a separate domain of life (Fig. 1A), their relationship to *Bacteria* and eukaryotes has been a matter of debate and is regarded to be of fundamental importance for our understanding of the origin of eukaryotes. Eukaryotic cells are highly compartmentalized and it has long been recognized that eukaryotic compartments, such as mitochondria (the site of ATP generation via oxidative phosphorylation) and chloroplasts (the organelles in which photosynthesis occurs in plants), evolved as a result of endosymbiosis, i.e., mitochondria and chloroplasts seem to be derived from Alphaproteobacteria and Cyanobacteria, respectively (e.g., reviewed in (*11*)). In contrast, the nature of the host cell taking up the progenitors of these compartments was unknown until recently; while some hypotheses suggested that this cell was a proto-eukaryote that already resembled extant eukaryotic cells, others point out that the host was an archaeon or even bacterium (*11–13*). For a long time, the prevailing view was that *Archaea* and eukaryotes represent two independent sister lineages in the tree of life (*2*, *14*), and it was unclear how the shared ancestor of *Archaea* and eukaryotes looked like. However, while certain phylogenetic analyses have supported

this model, others have suggested alternative scenarios, in which eukaryotes evolved from within the *Archaea* ((*15*) and reference therein).

Characteristic	Bacteria	Archaea	Eukarya
Membrane-enclosed nucleus	No	No	Yes
Chromosomal structure	Circular	Circular	Linear
Peptidoglycan in cell wall	Yes	No	No
Membrane lipids	Ester-linked	Ether-linked	Ester-linked
Glycerol	Glycerol-3-phosphate	Glycerol-1-phosphate	Glycerol-3-phosphate
Ribosomes (mass)	70S	70S	80S
Initiator tRNA	formylmethionine	methionine	methionine
Introns	No	No	Yes
Operons	Yes	Yes	No
RNA polymerase	One (4 subunits)	One (8-12 subunits)	Three (12-14 subunits)
Transcription factors required	No	Yes	Yes
TATA box in promoter	No	Yes	Yes

Table 1 Comparison of Selected Characteristics of the Major Domains of Life

The use of cultivation-independent genomic approaches combined with improved phylogenetic methods and more realistic evolutionary models have recently led new insights into the evolutionary history of the Archaea, their placement in the tree of life and eukaryogenesis. In particular, these analyses have provided increasing support for eukaryotes branching from within the *Archaea* (*10*, *15*) instead of as a sister lineage as originally assumed (Fig. 1 B–C). Though eukaryotes initially appeared to branch close to the TACK superphylum (discussed later in this chapter) (*16*) (Fig. 1B), it was challenging to pinpoint a specific archaeal lineage as being more closely related to eukaryotes than the others. The position of eukaryotes among *Archaea* became clearer with the recent discovery the Asgard archaea—a novel archaeal superphylum (*17*, *18*) (discussed later in this chapter). Phylogenomic analyses revealed that Asgard archaea form a sister group of eukaryotes (Fig. 1C) and harbor an extended set of proteins that were previously assumed to be specific to eukaryotes (*17*, *18*) (discussed later in this chapter). Together, these findings indicate that eukaryotes may have evolved from a symbiosis between an archaeal host cell and a bacterial endosymbiont, and also provide greater evidence in support of a two-domain tree of life (*19–22*), with *Archaea* and *Bacteria* representing two primary domains and eukaryotes being a secondary domain (*15*, *23*). Although the exact placement of eukaryotes with respect to the different members of the Asgard archaea remains to be elucidated, continued exploration of Asgard archaeal diversity will allow to further refine the position of Archaea and eukaryotes in the tree of life.

ARCHAEAL CELL BIOLOGY AND EUKARYOTIC SIGNATURE PROTEINS (ESPs)

In agreement with their close relationship to eukaryotes, *Archaea* encode informational processing machineries that closely resemble those of eukaryotic representatives. Although *Archaea* harbor a single circular chromosome like *Bacteria*, their replication machinery includes various components homologous (i.e., shared by common ancestry) to those of eukaryotes, while most functionally equivalent complexes in *Bacteria* are unrelated (*24*, *25*). For instance, *Archaea* and eukaryotes share homologous subunits comprising the origin of replication complex (ORC), a replicative helicase unit referred to as the CMG (Cdc45, MCM, GINS) complex, and the active replisome, which includes a two-subunit primase, a DNA polymerase sliding clamp and clamp loader, and DNA polymerases (*24*). Yet, some *Archaea* also encode components that are absent from both eukaryotes and *Bacteria* and others that are shared with *Bacteria*. For example the two-subunit DNA polymerase D (*24*, *26*) is unique to *Archaea* while the NAD+-dependent DNA ligase, the DNA gyrase, and the DNA primase DnaG are homologous to bacterial enzymes (*25*). In many cases, archaeal complexes seem to represent a simplified version of their counterparts in eukaryotes (*25*), the latter of which often encode additional paralogous enzymes (i.e., those that evolved by gene duplication), whose evolution involved sub-functionalization (*24*). For instance, while *Archaea* collectively encode three families of polymerase B, eukaryotes harbor the four polymerase B family enzymes referred to as Pol alpha, beta, gamma and delta (*24*, *26*). Notably, all of these eukaryotic enzymes seem to have evolved from two distinct archaeal polymerase B family homologs (*18*, *26*). Another interesting example represents the nucleosome: *Archaea* harbor histone-like proteins, which form a homodimeric histone complex in part homologous to the heterodimeric nucleosome of eukaryotes (*27*, *28*).

Archaeal transcription also shares several features in common with eukaryotes. While many archaeal genomes encode gene clusters reminiscent of bacterial operons, the archaeal transcription machinery represents a simplified version of their eukaryotic counterparts (*29*). For instance, the archaeal DNA-dependent RNA polymerase (RNAP) consists of 12-13 subunits, which are homologous to the subunits of the three eukaryotic RNA polymerases (RNAP I-III) (*29*). In contrast, RNAP of Bacteria consists of only five subunits, two of which are distantly related to archaeal RNAP subunits 1 and 2 (i.e., RpoA and RpoB). Transcription initiation, which is based on the same molecular mechanisms across the domains, also involves homologous transcription factors in *Archaea* and eukaryotes (*29*).

Similarities in the translational machinery between *Archaea* and eukaryotes are also evident. Archaeal ribosomes are of comparable size to bacterial ribosomes (70S), but share various ribosomal subunits uniquely with eukaryotes (*30*). Additionally, translation in *Archaea* is initiated by an initiator tRNA carrying methionine and several translation initiation factors, as is seen in eukaryotic organisms but contrasts with the use of formyl-methionine by bacteria. Further, a 22nd amino acid, pyrrolysine, has been identified uniquely in certain members of the *Archaea*, in particular methanogens (*31*).

Notably, *Archaea* not only share homologous replication, transcription, and translation machineries with eukaryotes, but have also been found to encode various so-called eukaryotic signature proteins (ESPs) (*32*), i.e., proteins that are generally absent from bacterial genomes while being central to the integrity and functioning of eukaryotic cells. These proteins include, for instance, components of the eukaryotic cytoskeleton (such as actin and tubulins), cell division and vesicle trafficking machineries, endosomal sorting complexes required for transport (ESCRT), as well as the proteasome and ubiquitin system (*10*).

In particular, members of the TACK archaea (discussed later in this chapter) including among others the *Cren-*, *Aig-* and *Thaumarchaeota* have early on been found to encode certain ESPs that were absent from *Euryarchaeota* (*15*, *16*, *33–35*). For instance, while *Euryarchaeota* use FtsZ as major cell division protein, many *Cren-* and *Thaumarchaeota* harbor a cell division system (also referred to as cdvABC system) that includes homologs of eukaryotic ESCRT-III and an ATPase related to vacuolar protein sorting-associated protein 4 (Vps4) (*36–39*). Furthermore, archaeal actin homologs referred to as crenactin, which are distantly related to eukaryotic actins, have been discovered in *Thermoproteales*, as well as in *Korarchaeota* (*40*). Yutin and coworkers identified distant homologs of eukaryotic tubulins—the artubulins—in the genomes of two species of *Thaumarchaeota*, "*Candidatus Nitrosoarchaeum limnia*" and "*Candidatus Nitrosoarchaeum koreensis*" (*41*), and an analysis of the "*Candidatus Caldiarchaeaum subterraneum*" composite genome revealed the presence of a presumably fully functional ubiquitin-like protein modifier system (*42*).

The discovery of the Asgard archaea (*17*, *18*), the currently most closely related archaeal sister lineage of eukaryotes, has recently revealed a variety of additional ESPs in *Archaea*. For instance, Asgard archaea not only encode additional homologs of eukaryotic informational processing machineries but also harbor simplified versions of the eukaryotic oligosaccharyl-transferase complex and ubiquitin modifier system. Furthermore, they encode an extended set of small GTPases (*17*, *18*), which are key regulators in eukaryotic cells with a central role in vesicle trafficking machineries (*43*). Additional central components homologous to eukaryotic vesicle transport and tethering were identified in the genomes of the *Thorarchaeota* (*18*). Further, Asgard archaea harbor protein domains homologous to the key domains of the three major eukaryotic ESCRT machinery complexes (ESCRT I-III) and a diversity of cytoskeleton-related proteins that are much more similar to their eukaryotic counterparts than those previously identified in *Archaea*. These include the lokiactins found across the Asgard representatives, as well as *bona fide* tubulins in *Odinarchaeota* (*10*, *17*, *18*). Notably, Asgard archaea also encode actin-regulating proteins, such as the profilins (*18*), which were recently shown to be functionally equivalent to those of eukaryotes (*44*).

Altogether, archaeal information processing machineries as well as an extended set of ESPs in members of the TACK and in particular the Asgard archaea, further testify to the archaeal origin of the eukaryotic cell. Importantly, the study of these complexes in *Archaea* can help to provide a better understanding of eukaryotic cell biology and provide insight into the relative timing of the evolution of cellular complexity.

ARCHAEAL CELL MEMBRANES AND CELLS WALLS

The composition of archaeal cell membranes differs fundamentally from those of *Bacteria* and eukaryotes (*45*). For instance, the glycerol used to make archaeal phospholipids is a stereoisomer of the glycerol used to build bacterial and eukaryotic membranes, i.e., while *Archaea* use glycerol-1-phosphate, eukaryotes and bacteria have glycerol-3-phosphate. Furthermore, *Archaea* harbor isoprenoid side chains instead of the fatty acid side chains found in Bacteria and eukaryotes. These isoprenoids are bound to the glycerol backbone by ether linkages contrasting with the ester linkages formed between the bacterial and eukaryotic glycerol and fatty acid moieties. Archaeal isoprenoid side chains in the two monolayers of the lipid bilayer can be linked, thereby giving rise to transmembrane phospholipids. The isoprenes can also form five-carbon ring structures, which may function in the stabilization of the membranes of archaeal species that live in high temperature environments. More than 100 different ether-type polar lipids, such as phospholipids and glycolipids, have been identified in *Archaea* (*46*).

Different archaeal representatives differ with regard to their cell walls. In contrast to *Bacteria*, *Archaea* lack peptidoglycan and are thus naturally resistant to antibiotics that impair the synthesis of peptidoglycan, such as penicillins. Some species of methanogenic *Archaea* contain cell walls of pseudopeptidoglycan (pseudomurein) that superficially resemble bacterial peptidoglycan but contain different components (e.g., N-acetyltalosaminuronic acid instead of and N-acetylmuramic acid) and have β-1,3 instead of β-1,4-glycosidic bonds. Yet, most archaeal species lack pseudomurein and instead harbor cell walls made of proteins, glycoproteins, or polysaccharides (*47*). For instance, a common cell wall structure found in *Archaea* is composed of a paracrystalline surface layer, termed S-layer, consisting of protein or glycoprotein moieties arranged in hexagonal patterns. Finally, some *Archaea*, such as certain members of the order *Thermoplasmatales*, lack cell walls altogether.

TAXONOMIC DIVERSITY OF ARCHAEA

The *Archaea* were originally divided into two major phyla, termed *Crenarchaeota* and *Euryarchaeota* (*2*). However, recent advances in culture-independent, high-throughput sequencing techniques have uncovered a large diversity of novel archaeal lineages, most of which remain uncultivated (*5*). Many of these newly discovered archaeal lineages are only distantly related to established lineages within the *Cren-* and *Euryarchaeota*, which has led to the proposal of many additional archaeal phyla and superphyla during the past years (*7*). Fig. 2 summarizes the current understanding of the archaeal phylogeny, including established and proposed phyla, classes, and orders, as well as their general physiological grouping and certain features discussed below. However, please note that there is currently no consensus on how to best classify archaeal lineages. Therefore, a widely accepted taxonomy of the *Archaea* remains to be established (*5*). In particular, there are currently two main classification schemes used: the classification suggested by Adam and coworkers that is implemented in NCBI (*7*) and the system introduced by the developers of the Genome Taxonomy Database (GTDB) (https:// gtdb.ecogenomic.org/). The latter of these was suggested to provide a standardized and rank-normalized genome-based classification system, which was recently used to revise the bacterial taxonomy (*48*).

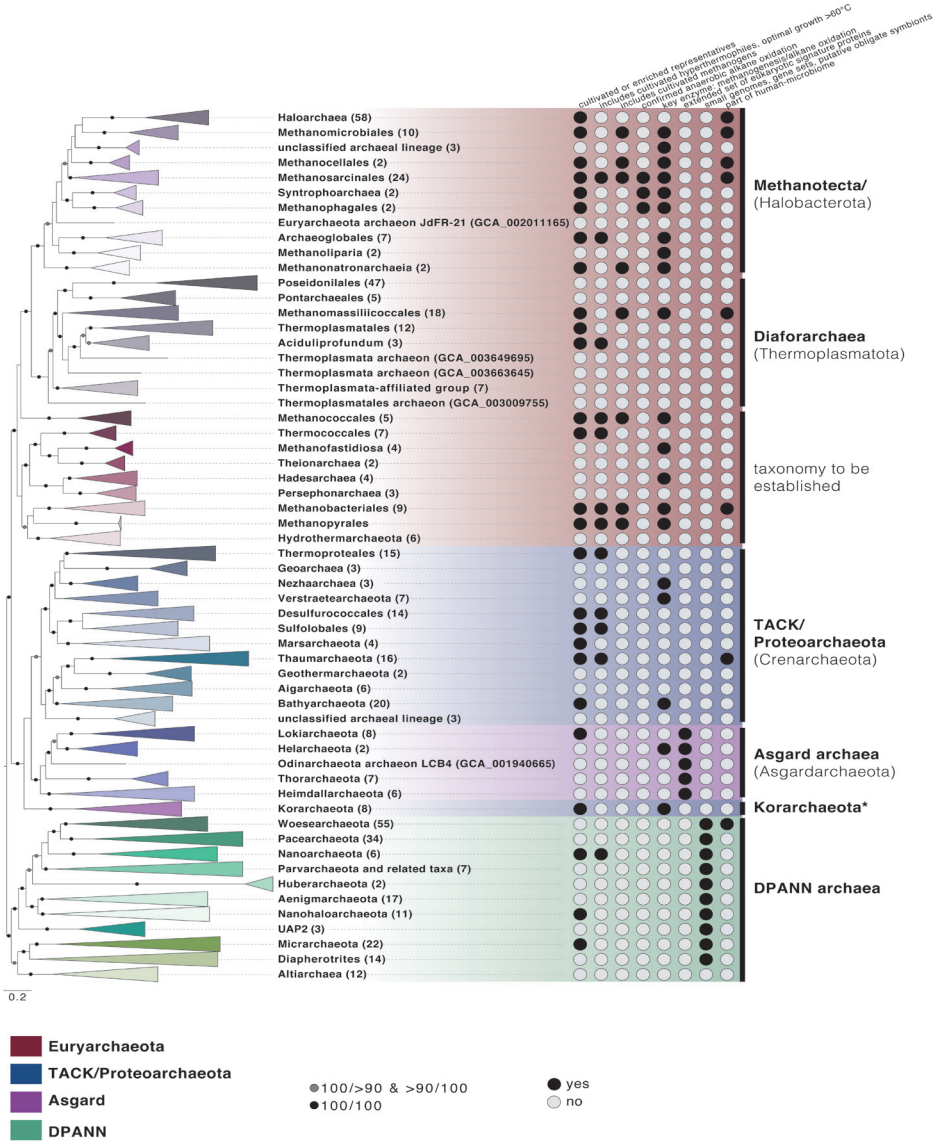


Fig. 2 | Depiction of the phylogenetic diversity of the *Archaea* and the presence/absence of key features. The unrooted phylogenetic tree was inferred with maximum likelihood using the IQ-tree software (with the C20+LG+F+R mixture model) and was based on an alignment of 11399 positions from a representative set of 569 archaeal taxa. Highly supported clusters (assessed by ultrafast bootstraps (*220*) and SH-like approximate-likelihood ratio test (*221*)) are indicated with gray or black dots based on their branch support values (see figure inlet). Taxa were predominantly named according to Adam et al. (*7*) (bold face), but alternative names suggested by GTDB (https://gtdb.ecogenomic.org/) are indicated in parenthesis when available. Numbers in brackets indicate the number of genomes/taxa per cluster. The presence and absence of certain features are shown with black and gray circles for each major taxonomic lineage (see figure inlet). Please note that the last column only reports those archaeal taxa that have been confirmed to be part of the human archaeome.

EURYARCHAEOTA

The *Euryarchaeota* (Fig. 2) comprise various cultivated and well-characterized archaeal species including the globally important methanogens (i.e., methane producers) as well as anaerobic methane-oxidizing Euryarchaeota (ANME) (*49*, *50*). Methanogens and ANME play a key role in the carbon cycle by anaerobically producing or consuming the potent climate gas, methane (*8*, *9*, *50–52*). However, research during the past years has shown that the *Euryarchaeota* are a phylogenetically and physiologically much more diverse radiation than originally thought (*5*, *7*). Indeed, it remains to be elucidated whether *Euryarchaeota* comprise a monophyletic group or phylogenetically distinct divisions, some of which may be more closely related to the TACK and Asgard archaea (*7*, *53*, *54*). In the following, we provide an overview of the major lineages comprising canonical and recently discovered lineages affiliating with the *Euryarchaeota*.

Methanotecta

The *Methanotecta* (Fig. 2), a recently proposed super-class (*7*), comprise the so-called class II methanogens (*Methanosarcinales*, *Methanomicrobiales*, *Methanocellales*), several phylogenetically distinct ANME archaeal lineages, the *Haloarchaeota*, *Archaeoglobales*, as well as the more recently described archaeal orders referred to as *Methanonatronarchaeia*, *Syntrophoarchaeales*, *Methanoliparales*, and *Methanophagales*. We present major features of these different lineages below.

Methanomicrobiales

The order *Methanomicrobiales* comprises several families, such as the *Methanocalculaceae*, *Methanoregulaceae*, *Methanospirillaceae*, *Methanomicrobiaceae*, and *Methanocorpusculaceae* (e.g., reviewed in (*55*)), and can be found in a variety of anoxic habitats, including wetlands, soil, oceans and freshwater, landfills, rice paddies, as well as associated with animals (*50*). Members of the *Methanomicrobiales* have diverse cell shapes, ranging from rods to cocci to plates, including motile and nonmotile species, and grow between 0°C and 60°C (*55*). Cells are often surrounded by glycoprotein-containing S-layers. Many *Methanomicrobiales* use hydrogen and carbon dioxide to form methane and all species are obligate anaerobes. They can use formate and alcohol but not acetate and methylated C1-compounds as substrates for methanogenesis, distinguishing them from the *Methanosarcinales* (*9*, *55*).

Methanosarcinales

The *Methanosarcinales* are closely related to the *Methano microbiales* and include families such as the *Methanosarcinaceae*, *Methanotrichaceae* (formerly *Methanosaetaceae*), and *Mether micoccaceae* (Table 1), as well as the *Methanoperedenaceae* (ANME-2d). While this order comprises diverse methanogenic organisms, it also includes representatives of the anaerobic methane-oxidizing *Euryarchaeota* ANME-2 and -3 lineages (*50*, *52*, *56*). Similar to the *Methanomicrobiales*, representatives of the *Methanosarcinales* are found in a range of anoxic habitats (*50*). Yet, in contrast to other methanogenic orders, the *Methanosarcinales* are known for their much wider substrate range for methanogenesis, i.e., members of this group not only use hydrogen and formate as substrates but also a variety of methylated compounds and

acetate (*8*, *9*). Considering that methanogenesis based on acetate may contribute up to two-thirds of methane released to the atmosphere, members of this group have important roles in the global carbon cycle (*8*, *51*). Representatives of the ANME-2 and -3 lineages use the reverse methanogenesis pathway to anaerobically oxidize methane (*52*). While some ANME-2 members can grow independently using nitrate, nitrite, or Fe(III) as electron acceptors (*57–59*), other ANME-2 grow in syntrophic consortia with bacterial partners (especially sulfate reducers) that serve as external electron sinks (*52*, *60*). Members of these groups are particularly abundant in the sulfate-methane transition zone in marine sediments and play an important role in the global carbon cycle by reoxidizing a large fraction of the methane produced in marine sediments before it can enter the atmosphere (*52*, *61*).

Methanophagales (ANME1)

The *Methanophagales* comprise another lineage of anaerobic methane-oxidizing archaea, also known as the ANME-1 lineage. While originally thought to affiliate with the *Methanosarcinales*, they were recently shown to represent a sister lineage of the *Syntrophoarchaeales* (*7*) (Fig. 2 and later in this chapter). Similar to the ANME-2 and -3 lineages that belong to the *Methanosarcinales*, members of this group occur in diverse marine, terrestrial, and freshwater environments (*62*), are particularly abundant in the sulfur-methane transition zone (*61*), and use the reverse methanogenesis pathway for the anaerobic oxidation of methane (AOM) (*63*). While ANME-1 has not been cultivated thus far, various lines of research have suggested that members are able to oxidize methane in syntrophy with sulfate-reducing bacteria (SRB) through direct electron transfer (*52*, *60*, *64*).

Methanocellales

Methanocellales represents a more recently described order of hydrogenotrophic methanogens that were originally referred to as Rice Cluster I (RC-I) (*65*) due to their initial discovery in rice paddy fields, where they are important producers of methane (*66*). The first representative of this order, *Methanocella paludicola*, was isolated from an anaerobic, propionate-containing enrichment culture (*65*) and represents a nonmotile anaerobe with rod-shaped cells thriving at temperatures between 25°C and 40°C (*65*). While the isolate performs methanogenesis using hydrogen, carbon dioxide, and formate, it uses acetate as a carbon source. Hydrogen is provided by its syntrophic partner, the bacterium *Syntrophobacter fumaroxidans* (*67*). Similar metabolic features were found in other representatives of this order, including *M. arvoryzae* (*68*) and *M. conradii* (*69*).

Syntrophoarchaeales

Syntrophoarchaeales (sometimes assigned to the *Methano sarcinales*; Table 1) represent a recently discovered group of anaerobic, alkane-oxidizing archaea usually found in hydrocarbon-rich sediments (*70*, *71*). For example, the first two representatives of this lineage, *Syntrophoarchaeum butanivorans* and *Syntrophoarchaeum caldarius*, were originally isolated from hydrothermal- and hydrocarbon-rich marine sediments of the Guaymas Basin (*71*, *72*). Notably, *Syntrophoarchaeales* grow by the anaerobic oxidation of butane as well as

propane, which are thought to be metabolized using the reverse methanogenesis pathway also operating in ANME archaea (*73*). In particular, they encode subunits homologous to the Methyl-Coenzyme M Reductase (MCR) complex, which represents the key enzyme of methanogens catalyzing the demethylation of CH3-S-CoM to methane (*51*). In *Syntrophoarchaeales*, MCR is thought to be used in reverse and to mediate the first step of the breakdown of short-chain alkanes eventually yielding carbon dioxide as an end product (*71*). As indicated by the names of members of this group, studied representatives grow syntrophically with the sulfate-reducing bacterium *Candidatus* Desulfofervidus auxilii.

Archaeoglobales

The *Archaeaoglobales* comprises species belonging to the genera *Archaeoglobus*, *Ferroglobus*, and *Geoglobus* (*74*). The *Archaeoglobus* sp. is believed to be predominantly composed of strictly anaerobic and hyperthermophilic members, growing optimally at 80°C and neutral pH. The best studied representatives are autotrophs and/or organotrophs and can reduce sulfate or sulfite during respiration (*75*). Species of *Ferroglobus* grow by oxidation of $Fe(II)S^{2-}$ and H_2 (*75*), whereas *Geoglobus* grows anaerobically in the presence of acetate and ferric iron (*74*). Recently, genomes of so far uncultivated members of the *Archaeoglobi* were reconstructed from environmental samples and shown to encode MCR-like protein complexes similar to those of methanogens and ANME archaea (*76*, *77*). Based on genomic inferences, it was suggested that the respective organisms may be able to grow by the oxidation of methane or alternative short-chain alkanes.

Methanoliparales

Methanoliparales is an uncultivated lineage within the *Methanotecta* that phylogenetically places between *Archaeoglobales* and a cluster comprising *Syntrophoarchaeales* and *Methanophagales*. *Methanoliparales* were first discovered in two metagenomes from a petroleum-enrichment culture and an oil seep and are represented by two metagenome-assembled genomes: *Candidatus Methanoliparum thermophilum* NM1a and *Candidatus Methanolliviera hydrocarbonicum* NM1b (*78*). Genomic analyses suggest that *Methanolipirales* are methanogens that encode the Wood-Ljungdahl carbon fixation pathway and are capable of beta-oxidation. Interestingly, both genomes code for two distinct MCR complexes, which may be involved in methanogenesis and the oxidation of alkanes, respectively.

Haloarchaeota

Halobacteria, herein referred to as *Haloarchaea*, are a diverse group of *Archaea*, most of which are adapted to high salinity. Salt requirements of these species range from 1.5 to 5.2 M NaCl, although most strains grow best between 3.5 and 4.5 M NaCl, at or near the saturation point of salt (36% w/v salts). In order to maintain osmolarity of their cells in high-salt environments, haloarchaeal members accumulate up to 5 M intracellular levels of KCl to counterbalance high extracellular salt concentration. As a result, the entire intracellular machinery, including enzymes and structural proteins, must be adapted to high salt levels. The proteins of all

haloarchaeal species have a very low isoelectric point and the genomes contain high GC contents that are well above 60% (*79*).

Some species of *Haloarchaea* are motile by means of tufts of flagella, although many species are nonmotile (*75*). *Haloarchaea* comprise various aerobic or facultative anaerobes and show diverse morphologies and shapes, including rods, cocci, and a multitude of pleomorphic forms (*75*, *80*). The lack of turgor pressure within haloarchaeal cells enables the cells to tolerate the formation of corners, and as such, some species are even triangular or square-shaped (*75*, *80*). Cell envelopes of coccoid *Haloarchaea* are stable in the absence of salt, while, noncoccoid species maintain their integrity only in the presence of high concentrations of NaCl or KCl (*75*). Non-coccoid species have a proteinaceous cell envelope with glycoprotein subunits forming a hexagonal pattern (*75*). Species of *Haloarchaea* are abundant in salt lakes, inland seas, and evaporating ponds of seawater, such as the Dead Sea and solar salterns. *Haloarchaea* often tint the water column and sediments in bright colors due to the presence of retinal-based pigments. Some of these pigments are capable of the light-mediated translocation of ions across cell membranes. The best known halobacterial pigment is bacteriorhodopsin, which is an outwardly directed proton pump. Bacteriorhodopsin is involved in energy conservation and is the only nonchlorophyll-mediated light energy transducing system known to date (*79*). Other retinal-based pigments found in *Haloarchaea* include halorhodopsin, which is an inward chloride pump involved in osmotic homeostasis, as well as sensory rhodopsin I and II (SRI and SRII, respectively). SRI and SRII can mediate positive and/or negative phototaxis (*79*).

Methanonatronarchaoeta

Another lineage of halophilic archaea are the *Methanonatronarchaeota*, which were first recovered from hypersaline anoxic lake sediments (*81*) and are currently represented by isolates from two distinct subgroups: the soda lake isolate *Methanonatronarchaeum thermophilum* AMET and the salt lake isolate *Candidatus Methanohalarchaeum thermophilum* HMET (*81*). *Cand. M. thermophilum* has motile, coccoid cells that are around 0.4 µM in diameter and are surrounded by an S-layer. These anaerobic organisms tolerate a range of pH (between 6.5 and 8 [HMET] and 9.5 and 9.8 [AMET]) and grow optimally at a temperature of 50°C and salt concentrations of 4 M by accumulating high concentrations of potassium inside their cells for osmotic balance ("salt-in strategy") (*81*, *82*). *Cand. M. thermophilum* is a heterotrophic methanogen that grows with C1-methylated compounds as electron acceptors, such as methanol or trimethylamine, and formate or hydrogen as electron donors (*81*). The 16S rRNA gene analyses indicate that *Methanonatronarchaeota* are the first cultured representatives of the SA1 group, which is commonly found in hypersaline environments (*81*, *83*). Yet, the exact placement of *Methanonatronarchaeota* in the archaeal tree of life is still debated. While initial phylogenetic analyses placed this lineage sister to *Haloarchaea* (*81*), recent analyses have suggested that the *Methanonatronarchaeota* form an early diverging lineage of the *Methanotecta* (*84*).

DIAFORARCHAEA

The *Diaforarchaea* comprise a recently suggested superclass (*7*) that includes the *Thermoplasmata* and related lineages, such as the diverse and abundant Marine Group II and III archaea (*85*, *86*), now also known as the *Poseidoniales* and *Pontarchaeales*, respectively, as well as a recently discovered new order of methanogens, the *Methanomassiliicoccales* (*87*).

Thermoplasmatales

The *Thermoplasmatales* comprise the genera *Acidiplasma, Thermoplasma, Picrophilus, Cuniculiplasma,* and *Ferroplasma. Cunicuplasma, Thermoplasma,* and *Ferroplasma* are the only cultivated archaeal representatives that lack cell walls (*75*, *88*). Species of *Thermoplasma* are facultative anaerobes and obligate heterotrophs, using elemental sulfur for respiration. Most members of this group are thermoacidophiles and grow optimally at 60°C and pH 2 (*75*). For instance, representatives may be found in self-heating coal refuse piles and in acidic solfatara fields (*75*).

Members of the *Picrophilus* are the most acidophilic organisms known so far (*89*). They form irregular cocci that are 1–1.5 µm in diameter and contain S-layer cell walls (*75*). *Picrophilus* are thermophilic and hyperacidophilic and grow at temperatures between 47°C and 60°C and pH ranges of 0–3.5 (*75*). Their ability to grow at pH values near 0 and at high temperatures has shifted the physicochemical boundaries at which life was considered to exist.

In contrast to other members of the *Thermoplasmatales*, *Ferroplasma* are not thermophilic and can grow autotrophically using ferrous iron as energy and inorganic carbon as a carbon source. Representatives can be found in a variety of acidic environments with stable chemical conditions, such as ore deposits, mines, and acid mine drainage systems (natural or man-made), as well as in areas with geothermal activity (*90*, *91*). Representatives of this family are cell wall-lacking extreme and obligate acidophiles that are able to grow at pH values around 0. Together with members of *Picrophilum*, they comprise a group of the most extreme acidophilic organisms known, members of which tolerate high concentrations of iron, copper, zinc, and other metals (*91*).

Aciduliprofundales

Aciduliprofundales, formerly named the “deep-sea hydrothermal vent euryarchaeota 2” (DHVE2) lineage, is currently represented by the cultivated *Aciduliprofundum boonei* (*92*, *93*). As the original name suggests, *Aciduliprofundales* are predominantly found across hydrothermal vents, where they can represent up to 15% of the archaeal community (*92–94*). *A. boonei* is an anaerobic heterotroph that ferments peptides and is able to reduce elemental sulfur or ferric iron at a pH between 3.3 and 5.8 (optimum pH 4.6) and an optimal growth temperature of 70°C (*92*). This organism is motile with a single flagellum and has pleomorphic cells of about 0.6–1 µM in diameter that are surrounded by a single S-layer.

Methanomassiliicoccales

The order *Methanomassiliicoccales* represents the first lineage of the *Thermoplasmata* known to comprise methanogenic members (*87*), several of which have been isolated, such as *Methanomassiliicoccus luminyensis* (*95*, *96*), *Candidatus Methanomethylophilus alvus* (*97*), and *Candidatus Methano plasma termitum* (*98*). *Methanomassiliicoccales* are widely distributed in wetlands and sediments as well as the gastrointestinal tracts of animals including those of humans and cows (*87*, *99*, *100*). Members of this group comprise H_2-dependent methylotrophic methanogens, which are able to use methylated amines (*100*) including mono-, di-, and trimethylamines for methanogenesis. Considering that the latter compounds have been implicated in human disease, gut-associated members of the *Methanomassiliicoccales* may play an important role in human health (*100*).

Poseidoniales

The *Poseidoniales* (*101*), formerly Marine Group II (MG II), lack any cultured representatives and are mainly known from 16S rRNA gene diversity assays and genomic analyses. *Poseidoniales* are often found in the photic zone of marine waters and can present up to 15% of archaeal cells in the Atlantic ocean (*102–104*). They are further divided into *Candidatus Poseidonaceae* (MGIIa) and *Candidatus Thalassarchaeacea* (MGIIb), whose abundances seasonally fluctuate, i.e., members of MGIIa and MGIIb are more abundant in the summer and winter, respectively (*105*). Members of this group comprise aerobic heterotrophs with the potential to utilize a range of substrates such as proteins, peptides, amino acids, fatty acids, carbohydrates, xenobiotics, and agar (*101*, *106–110*). In addition, some representatives of the class *Ca. Poseidoniia*, found in the photic zone, encode proteorhodopsin indicative of a photoheterotrophic lifestyle (*101*, *107*, *110*).

Pontarchaeales

The order *Pontarchaeales*, or Marine Group III, are often found in the deep ocean, while being less abundant in the photic zone (*102*, *111*). Based on genomic data, it was inferred that deepsea *Pontarchaeales* likely represent motile heterotrophs that might degrade proteins, carbohydrates, and lipids (*112*). In contrast, surface dwelling members of the *Pontarchaeales* seem to encode photolyase and rhodopsin genes and in turn may be photoheterotrophs (*111*). Notably, both the *Pontarchaeales* and the *Poseidoniales* lack the key archaeal lipid biosynthesis gene encoding glycerol-1 phosphate dehydrogenase, such that it is currently unclear whether members of these orders encode canonical archaeal lipids (*45*). In particular, the presence of genes for glycerol-3 phosphate dehydrogenase, which is essential in the synthesis of bacterial lipids, has led to the suggestion that these *Archaea* may have mixed membranes (*45*, *101*).

OTHER EURYARCHAEOTA

The following section provides an overview of additional lineages affiliating with the *Euryarchaeota*, including methanogenic lineages that have been extensively studied in the past. However, some analyses indicate that at least some of these orders may be more closely related to the TACK and Asgard archaea (*7*, *53*, *54*).

Methanococcales

As the name implies, the *Methanococcales* include representatives with coccoid shapes and proteinous cell walls (*75*). All members of this lineage are thought to be strict anaerobes that obtain energy by the reduction of CO2 to methane (*9*) and comprise mesophilic (e.g., *Methanococcus*) to thermophilic (e.g., *Methanothermococcus*) to hyperthermophilic (e.g., *Methanocaldococcus*) taxa (*75*).

Thermococcales

Members of the *Thermococcales* represent anaerobic heterotrophs that utilize a wide range of organic compounds, including amino acids, a variety of sugars, and organic acids such as pyruvate. When available, they can use elemental sulfur as the terminal electron acceptor. Extensive research has been carried out on the metabolism of cultivated representatives and led to the discovery of unique enzymes and pathways (*113*). Certain members of the *Thermococcales* represent important model organisms. For example, the hyperthermophilic *Pyrococcus furiosus*, which grows anaerobically at temperatures near 100°C using carbohydrates and peptides as carbon and energy sources (*75*), has been extensively used to study thermostable enzymes and adaptations to high-temperature environments (*114*).

Methanobacteriales

The *Methanobacteriales* comprise another lineage of methanogenic archaea that reduce CO2 or methyl compounds with H_2, formate, or secondary alcohols as electron donors. They include rod-shaped, lancet-shaped, or coccoid members, which contain cell walls made of pseudopeptidoglycan. *Methanobacteriales* are widely distributed in nature and are found in anaerobic habitats such as aquatic sediments, soil, anaerobic sewage digesters, and the gastrointestinal tracts of animals to name a few (*50*, *75*).

Methanopyrales

The *Methanopyrales* consists of a single genus, *Methanopyrus*, comprising rod-shaped members with cell walls made of pseudopeptidoglycan (*75*). Known *Methanopyrus* are hyperthermophilic, and grow between 84°C and 110°C, with optimal growth at 98°C. Similar to other methanogenic lineages, members of this group have a chemolitoautotrophic lifestyle converting CO_2 and H_2 to methane (*9*, *75*). While it has proven difficult to resolve the exact phylogenetic placement of the *Methanopyrales* relative to other archaea, it has recently been suggested that this lineage forms a monophyletic clade together with the *Methanobacteriales* and the *Methanococcales* referred to as *Methanomada* (*7*). However, it remains to be determined whether these so-called group 1 methanogens (*9*) are indeed closely related phylogenetically (Fig. 2).

Methanofastidiosales

Methanofastidiosales represent a recently discovered and thus far uncultivated archaeal lineage (also known as WSA2 or Arc I), whose members are present in diverse environments including sediments, groundwater, and bioreactors (*115–117*). Metagenomic approaches have

enabled the reconstruction of genomes of representatives of the *Methanofastidiosales* from wastewater-treatment bioreactors (*117*). While members of this group encode key genes for methanogenesis, they lack genes related to carbon-fixation pathways and were suggested to solely use methylated thiols as substrates for methanogenesis (*117*).

Theionarchaeota

Theionarchaea (formerly Z7ME43) represents another clade of uncultivated archaea, which forms a sister lineage of the *Hadesarchaea* (see next) and was originally discovered in water-filled limestone sinkholes in northeastern Mexico (*118*). This clade is currently represented by two genomes that were recovered from the White Oak River Estuary in North Carolina (*119*). Genomic analyses indicated that *Theionarchaea* might conserve energy by peptide fermentation.

Hadesarchaea

The *Hadesarchaea*, which were originally referred to as the South-African Gold Mine Miscellaneous Euryarchaeal Group (SAGMEG), are distributed in a variety of anoxic environments, including the terrestrial subsurface as well as marine sediments, which cover a wide span of temperatures (*120–123*). The first genomes of members of this clade were reconstructed from the water column of the White Oak River estuary (*123*) as well as Yellowstone National Park (YNP) hot spring sediments and indicated the capability of anaerobic CO oxidation potentially coupled to nitrite or H_2O reduction (*123*). Notably, another genome of a member of the *Hadesarchaea* was recently obtained from a hot spring metagenome and shown to encode a *mcr*-like operon. Based on phylogenetic analyses of MCR subunits as well as genomic analyses, it was suggested that these *Hadesarchaea* may represent alkane-oxidizing archaea similar to members of the *Syntrophoarchaeales* (*124*) and perhaps some representatives of the *Bathyarchaeota* (*50*).

Persephonarchaea

The Mediterranean Sea Brine Lakes 1 (MSBL1) clade, now referred to as the *Persephonarchaea* (*7*), is another lineage of uncultivated archaea that is closely related to the *Hadesarchaea*. The *Persephonarchaea* are commonly found in marine hypersaline environments (*125*, *126*) and comprise potential anaerobic mixotrophs that may conserve energy through sugar fermentation but may also be able to fix inorganic carbon (*127*). Genomic inferences suggest that MSBL1 archaea synthesize trehalose as putative osmolyte to encounter the high salt conditions in their environment (*127*).

Hydrothermarchaeota

The *Hydrothermarchaeota* (*7*), also known as the Marine Benthic Group-E (MBG-E), were originally discovered in marine deep-sea sediments (*128*) and represent an uncultivated archaeal lineage widely distributed in deep subseafloor environments. Genomes from members of this group have been reconstructed from metagenomes of the Juan de Fuca Ridge flank, Guaymas Basin hydrothermal sediments, and the Mid-Atlantic Ridge of the South Atlantic Ocean (*129–131*). Genomic analyses have indicated that *Hydrothermarchaea*

are metabolically versatile (*131*) and include putative anaerobic chemolithoautotrophs that use carbon monoxide and/or hydrogen as electron donors as well as a variety of electron acceptors including nitrate and sulfate (*132*, *133*).

THE TACK SUPERPHYLUM

The TACK superphylum was originally introduced to describe the *Crenarchaeota* and the related phyla referred to as the *Thaumarchaeota, Aigarchaeota*, and *Korarchaeota* (*16*). During the past years, many additional lineages affiliating with the TACK archaea have been discovered through metagenomics and single cell genomics approaches and the TACK lineage has therefore been suggested to be referred to as the *Proteoarchaeota* (*134*). However, a consensus has yet to be reached regarding both the naming as well as the validity of using a superphylum as a taxonomic level. In the following sections, we introduce canonical and recently discovered clades belonging to the TACK archaea.

Crenarchaeota

The *Crenarchaeota* includes a diversity of (hyper-) thermophilic archaeal species, many of which have been discovered through cultivation-based approaches before the onset of the genomics era in microbiology and now represent important model organisms. This taxon is composed of a single class, the *Thermoprotei*, which is subdivided into three to five subclades, the *Thermoproteales, Sulfolobales, Desulfurcoccales* as well as the *Fervidicoccales* and *Acidilobales*. However, the latter two may in fact belong to the *Desulfurococcales* (Fig. 2). Cultured crenarchaeal species are morphologically diverse, and include rods, cocci, filamentous, and disk-shaped cells. Almost all cultured species are obligate (hyper-) thermophiles, with optimal growth temperatures ranging from 70°C to 113°C and many members are also acidophiles and capable of metabolizing sulfur. Representatives of the *Crenarchaeota* thrive in environments such as hot solfataras, volcanic areas, as well as hydrothermal vents at the bottom of the ocean. A variety of metabolic capabilities have been described in the different members of the *Crenarchaeota*. For instance, some *Thermoproteales* are chemolithoautotrophs, using carbon dioxide as a carbon source and conserving energy by the conversion of hydrogen and elemental sulfur to hydrogen sulfide. Others respire various organic substrates using oxygen, sulfur, nitrate, or nitrite as electron acceptors (*75*). Many members of the *Desulfurococcales* are strict anaerobes and neutrophiles to weak acidophiles, growing optimally at pH 5.5–7.5 (*135*). Representatives of the *Sulfolobales* are acidophilic hyperthermophiles, which can grow lithoautotrophically by oxidizing sulfur or chemoheterotrophically on simple reduced carbon compounds using sulfur derivatives as electron acceptors. Notably, the *Crenarchaeota* include several members that have been shown to be hosts of the small-celled *Nanoarchaeota* (*136–140*) (see later in this chapter). In particular, the biocoenosis between *Ignicoccus hospitalis*, a member of the *Desulfurococcales*, and its nanoarchaeal ectosymbiont, *Nanoarchaeum equitans*, has been extensively studied and provides important insights into archaeal cell biology and cell-cell communication (*141*). For instance, investigation of *I. hospitalis* has revealed remarkable cellular features including the presence of two outer membranes surrounding a large periplasmic space as well as an endomembrane system reminiscent of eukaryotic cells (*142*).

Thaumarchaeota

Environmental 16S rRNA-based surveys in the early 1990s have led to the discovery of uncultivated archaeal lineages distantly related to the *Crenarchaeota* in moderate marine and terrestrial ecosystems. The subsequent cultivation of the first representatives of these so-called mesophilic *Crenarchaeota* (also MG1) from marine and terrestrial environments (*143*) and the study of the first genomes of members of this group (*144*, *145*), revealed that they form a separate phylum within the *Archaea* referred to as the *Thaumarchaeota* that distantly affiliates with the *Crenarchaeota*. Most cultivated *Thaumarchaeota* are chemolithoautotrophic ammonia-oxidizing archaea (AOA), which play an important role in the nitrogen and carbon cycles in both aquatic and terrestrial environments (*146*). However, the reconstruction of genomes of deep-branching *Thaumarchaeota* has recently led to the suggestion that not all members of this group are AOA but instead represent chemoorganotrophs that may reduce oxygen, nitrate, or sulfur (*147*). This notion was recently confirmed with the isolation of the thermoacidophilic, sulfur- and iron-reducing organoheterotrophic *Conexivisphaera calidus*, a potentially early diverging member of the *Thaumarchaeota* (*148*).

Aigarchaeota

The *Aigarchaeota* represent another proposed candidate phylum that comprises species of the Hot Water Crenarchaeotic Group I (HWCGI), members of which have not been cultivated so far. Genomic analyses of the first representatives of this group have suggested that the *Aigarchaeota* comprise both facultative and obligate anaerobes, which may respire a variety of organic substrates and perhaps also hydrogen and carbon monoxide using oxygen or oxidized sulfur or nitrogen compounds as electron acceptors (*42*, *149–152*). Furthermore, several representatives seem to have the ability to fix inorganic carbon. *Aigarchaeota* seem to predominantly inhabit thermally heated terrestrial and marine ecosystems, including hot springs, subsurface aquifers, and mine fracture waters (*150*, *152*).

Korarchaeota

The *Korarchaeota* comprises a group of uncultivated *Archaea* that had already been discovered in the late 1990s in terrestrial and marine thermal environments (*153*). The first member of this clade, referred to as "*Ca. Korarchaeum cryptofilum*," was shown to comprise ultra-thin, needle-shaped cells measuring up to 100 µm in length. Genomic analyses indicated that this organism represents a peptide fermenter with a unique set of informational processing genes, which early on indicated that it comprises the first member of a distinct archaeal phylum (*154*). Recently, genomes of additional members of the *Korarchaeota* have been recovered from deep-sea hydrothermal vent sediments (*130*) and hot spring environments (*18*, *124*, *155*) providing novel insights into the metabolic features of this clade. Notably, genomic analyses revealed that certain members of the *Korarchaeota* harbor the key genes for methanogenesis, (*155*) which may for instance enable methanogenesis from methanol and hydrogen or the coupling of the anaerobic oxidation of methane with sulfite reduction (*155*).

Bathyarchaeota

Bathyarchaeota were originally discovered through 16S rRNA gene surveys in hot springs (*153*) and were referred to as Miscellaneous Crenarchaeota Group (MCG) (*156*) due to their distant affiliation with cultivated *Crenarchaeota*. This extremely diverse phylum is now subdivided into at least 25 subgroups, which are defined at family and order level (*157*). Notably, members of this putative phylum-level lineage can be found in a diversity of anoxic marine, terrestrial, and hydrothermal environments including marine sediments and often represent the most abundant archaeal community members (*157–159*). Based on genomic analyses, it is inferred that many *Bathyarchaeota* are heterotrophs with a wide substrate range including acetate, proteins, and aromatic compounds such as lignin (*157*, *160*). However, the *Bathyarchaeota* also includes putative acetogenic species (*161*) as well as organisms with *mcr* genes (*162*), which are closely related to those of *Syntrophoarchaea* (*71*). In turn, it has been suggested that some members of the *Bathyarchaeota* may be able to mediate the anaerobic oxidation of short-chain alkanes (*50*).

Geoarchaeota

Geoarchaeota, also Novel Archaeal Group 1 (NAG1), are often found in hypoxic to oxic, hot, acidic, iron-rich springs (*163–165*) and represent a lineage of thus far uncultivated archaea which seem to be closely related to or part of the Crenarchaeota (*164*, *166*). Based on genomic inferences, it has been suggested that the *Geoarchaeota* are likely motile and might conserve energy through the oxidation of carbon monoxide, peptides, and/or carbohydrates using oxygen as a terminal electron acceptor (*149*, *164*).

Verstraetearchaeota

Verstraetearchaeota were originally discovered in deep South-African Gold mine microbial communities through 16S rRNA gene surveys and were referred to as Terrestrial Miscellaneous Crenarchaeota Group (TMCG) (*120*). Members of this group seem to be widely distributed and are also found in hydrocarbon-rich environments, sediments, soil, and wetlands (*167*). First insights into the metabolic features of members of this group were derived from genomes assembled from anoxic digesters, named *Methanomethylicus sp.* and *Methanosuratus sp.* (*167*). Subsequently, additional representatives were discovered and referred to as *Methanohydrogenales* and *Methanomediales* (*168*). Notably, the *Verstraetearchaeota* comprise members with *mcr*-gene operons most similar to those found in methanogenic *Euryarchaeota*. In turn, based on genomic inferences, it was suggested that the *Verstraetearchaeota* likely include anaerobic methylotrophic as well as hydrogenotrophic methanogens (*167–169*).

Nezhaarchaeota

Nezhaarchaeota are a recent addition to the TACK superphylum represented solely by uncultivated members, whose genomes were assembled from hot spring metagenomes and hydrothermal sediments (*77*). Notably, the *Nezhaarchaeota* encode a MCR protein cluster and are potential hydrogenotrophic methanogens (*77*).

Marsarchaeota

The *Marsarchaeota*, or "Novel Archaeal Group 2" (NAG2), are typically found in geothermal, iron oxide-rich mats (*163*). The first genomes of members of this lineage were recently recovered from thermal (50–80°C) and acidic (pH 2.5–2.5) microbial mats from Yellowstone National Park (*170*) and led to the suggestions that the *Marsarchaeota* are aerobic chemoorganotrophs that degrade lipids, peptides, and carbohydrates and may be able to reduce ferric oxide.

Geothermarchaeota

The *Geothermarchaeota* represents one of the most recent additions to the *Archaea* and is thus far only represented by uncultivated members, whose genomes have been reconstructed from metagenomes from the Juan de Fuca Ridge subseafloor (*129*) and hydrothermal vent sediments in the Guaymas Basin (*130*). Little is yet known about the lifestyle and ecological roles of *Geothermarchaeota*, and in-depth genomic analyses will be necessary to infer their metabolic potential.

THE ASGARD SUPERPHYLUM

The Asgard superphylum is a recently described archaeal radiation, which comprises several different archaeal clades of high taxonomic rank (likely phylum-level) (*17*, *18*). Notably, phylogenetic and comparative genomic analyses have indicated that this archaeal clade includes the closest archaeal sister lineage of eukaryotes (discussed previously in this chapter). Members of this superphylum have originally been discovered in sediments all around the world, in which they can comprise a significant fraction of the microbial diversity. In the following, we briefly introduce the major metabolic features of the currently known members of the Asgard archaea, i.e., the *Loki-*, *Thor-*, *Odin-*, *Hel-*, and *Heimdallarchaeota*.

Lokiarchaeota

The *Lokiarchaeota* represents an archaeal lineage originally referred to as the Deep Sea Archaeal Group (DSAG) or Marine Benthic Group B (MBGB) archaea, which are abundant in diverse marine sediments (*94*, *171*). For example, the *Lokiarchaeota* comprise up to 10% of the microbial community in cold sediments near Loki's Castle hydrothermal vent field from which the first metagenomes were obtained (*17*, *18*). Members of the *Lokiarchaeota* might be autotrophs using the Wood-Ljungdahl pathway for carbon fixation (*172*). However, genomic analyses also revealed the potential for the use of a variety of organic carbon substrates, suggesting that representatives of the *Lokiarchaeota* may predominantly rely on fermentative growth (*20*). In fact, the successful cultivation of the first *Lokiarchaeote*, *Candidatus Prometheoarchaeum syntrophicum*, revealed that this organism ferments amino acids enabled through a syntrophic interaction with hydrogen- or formate-consuming partner organisms (*22*).

Thorarchaeota

The *Thorarchaeota* share many metabolic features with the *Lokiarchaeota* (*20*, *173*, *174*). Currently known representatives harbor a variety of genes likely encoding proteins involved in the usage of organic substrates. Furthermore, they encode the Wood-Ljungdahl pathway,

which could be used for carbon fixation or serve as an electron sink during growth on organics. In contrast to currently known *Lokiarchaeota*, members of this group also harbor a putative NADH dehydrogenase that may enable respiratory growth in addition to fermentation (*20*). Based on current environmental survey data, the *Thorarchaeota* seem less abundant than the *Lokiarchaeota* but occur in a wide variety of anoxic environments (*18*).

Heimdallarchaeota

Thus far known representatives of the *Heimdallarchaeota* are metabolically diverse and differ from other Asgard lineages (*20*). While genomic analyses indicate that they are able to utilize a large variety of organic substrates similar to other members of the Asgards, they do not seem to be fermentative organisms and current members lack the Wood-Ljungdahl pathway. Instead, they encode a membrane-bound electron chain, which allows growth using oxygen and nitrite as electron acceptors (*20*, *21*). *Heimdallarchaeota* are currently thought to comprise the archaeal lineage most closely related to the archaeal ancestor of eukaryotes. However, though found in a variety of environmental samples including anoxic sediments and oxygenated waters, they are generally less abundant than the *Loki-* and *Thorarchaeota*.

Odinarchaeota

Odinarchaeota are currently represented by a single genome, which was obtained from a hot spring metagenome (*18*). Similar to other members of the Asgard superphylum, *Odinarchaeum* encodes the ability to use organic compounds as growth substrates (*20*). Yet, it lacks the key enzyme of the Wood-Ljungdahl pathway and instead encodes membrane-bound hydrogenases, which suggests that the thermophilic *Odinarchaeum* may conserve energy through fermentation of organic substrates to hydrogen, acetate, and carbon dioxide. Members of the *Odinarchaeota* are thought to predominantly inhabit thermal environments such as hot spring sediments and hydrothermal vents (*18*).

Helarchaeota

The *Helarchaeota* represent the most recently discovered clade within the Asgard archaea (*175*). While they harbor similar gene sets as the *Loki-* and *Thorarchaeota*, currently known representatives of this lineage also contain *mcr*-gene clusters. Phylogenetic analyses of the encoded proteins revealed their close relationship with proteins of *Syntrophoarchaea* opening the possibility that certain members of the Asgard archaea have the potential to anaerobically oxidize short-chain alkanes, perhaps in syntrophy with microbial partners (*175*). However, the environmental distribution of the *Helarchaeota* and the functional importance of this potential alkane metabolism in Asgard archaea remain to be determined.

THE DPANN SUPERPHYLUM

The DPANN superphylum is the fourth major radiation in the *Archaea*, besides the *Euryarchaeota*, TACK, and Asgard archaea (*149*, *176*, *177*). Currently, this radiation is assumed to comprise a large diversity of distinct archaeal clades most of which seem to predominantly include members with extremely small genomes and cell sizes that are thought to depend

on partner organisms for growth and survival (*177*). While first defined in reference to the *Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota,* and *Nanohaloarchaeota* (DPANN) (*149*), additional lineages such as the *Micrarchaeota, Pacearchaeota, Woesearchaeota*, and *Huberarchaeota* are now also considered members of this group (*177*, *178*). Furthermore, the *Altiarchaeota* (*179*), representatives of which do not have reduced genomes, are sometimes considered to belong to the DPANN (*177*, *180*). However, the boundaries between certain clades of DPANN and other archaea (in particular the *Euryarchaeota*) are not well defined and it remains to be established which lineages indeed belong to a monophyletic (i.e., sharing a common ancestor) DPANN clade (*180*).

Nanoarchaeota

The first representative lineage of the DPANN archaea was already discovered in 2002, i.e., long before the DPANN radiation was known. In particular, Huber and coworkers discovered a small-celled organism in cultures of the crenarchaeaum, *I. hospitalis,* which they referred to as *N. equitans* (*136*). Initially, it was suggested that this organism is the first representative of a novel phylum called *Nanoarchaeota* or may represent a highly derived member of the *Euryarchaeota* (*136*, *181*). However, upon the genomics-based discovery of additional archaeal lineages represented by organisms with small genomes, which affiliated with *Nanoarchaeota*, it was proposed that the *Nanoarchaeota* belong to the DPANN radiation (*149*).

Notably, the nanosized hyperthermophilic *N. equitans* is obligately host-dependent and grows as an ectoparasite on the surface of *I. hospitalis* (*182*, *183*). It lacks genes for many major metabolic pathways and in turn depends on its host for the acquisition of diverse metabolites likely including lipids, amino acids, and ATP. In line with this, the genome of *N. equitans* represents one of the smallest known genomes of any extracellular organism (480 kb) (*184*). However, compared to the genomes of many bacterial endosymbionts, the genome of *N. equitans* does not show evidence of pseudogenes and contains a full complement of tightly packed genes encoding informational processing machineries (*184*). Similar trends have recently been seen in other representatives of the DPANN radiation (*177*). Members of the *Nanoarchaeota* have been found in a variety of thermal environments ranging from hydrothermal vents to hot springs and are now assumed to infect a variety of different crenarchaeal hosts. For instance, additional *Nanoarchaeaota*, such as *Candidatus Nanobsidianus stetteri*, *Candidatus Nanopusillus acidilobi*, and *Candidatus Nanoclepta minutus* have recently been successfully co-cultivated with their crenarchaeal hosts referred to as *Acidolobus sp., Acidilobus sp.* 7A, and *Zestosphaera tikiterensis*, respectively (*138*, *185*, *186*).

Overview of Other Putative DPANN Clades

Most of the other DPANN clades are represented by genomes reconstructed through metagenomics or single-cell genomics approaches. However, recent cultivation efforts have led to the enrichment of the first co-culture of a member of the *Nanohaloarchaeota* with its haloarchaeal host, i.e., *Ca. Nanohaloarchaeum antarcticus* and *Halorubrum lacusprofundi* (*187*), and of members of the *Micraarchaeota* members with their putative archaeal partners

belonging to the *Thermoplasmatales* (*188*, *189*). Even though *Ca. N. antarcticus* has a larger genome and metabolic gene repertoire than *N. equitans*, it seems to obligately depend on its host for growth and survival (*187*).

Additional insights into the diversity and metabolic potential of members of the various DPANN clades predominantly derive from genomic inferences (*6*, *177*, *180*). For instance, the *Woese-* and *Pacearchaeota* seemingly represent extremely diverse DPANN lineages whose members differ in the extent of genome reduction and metabolic capabilities. However, all representatives lack major and essential metabolic pathways indicating obligate host dependency. Few representatives of the DPANN, such as members of the *Diapherotrites* (*190*), may be able to conserve energy through fermentation. But the lack of some biosynthetic pathways indicates that they still depend on the external acquisition of certain amino acids, vitamins and/or cofactors (*177*, *180*).

While much has to be learned about the DPANN archaea, the discovery of this large diversity of putative archaeal symbionts and the occurrence of certain representatives in almost all environments on Earth indicates that the future investigation of this radiation will be crucial for our understanding of both the evolution and ecology of *Archaea* and their impact on global nutrient cycles.

Altiarchaeota and its Symbiont—A Member of the Huberarchaeota

The *Altiarchaeota* represent a lineage variably affiliating either with the DPANN archaea or *Euryarchaeota* (*6*, *135*, *177*, *179*, *180*, *191*) in phylogenetic analyses depending on the type of analysis (e.g., with regard to model choice) and data used. *Altiarchaeota* (formerly also referred to as SM1 *Euryarchaeota*) were originally discovered in a cold (~10°C), sulfurous Moor in Germany (*135*) but can also be found in sulfidic springs (*192*, *193*), marine sediments, hot springs, and aquifers (*191*). Notably, some members of the *Altiarchaeota* are found in microbial consortia that display a unique morphology described as a **"**string-of-pearls," which is several millimeters in length and consists of tiny white pearls (0.5–3 mm diameter) connected by thin threads (*135*). The outer part of the pearl is composed of bacteria, such as the *Gammaproteobacterium Thiotrix uunzi* (*194*) [194] or the *Epsilonproteobacterium* IMB1 (*195*), while the inside is dominated by *Altiarchaeota* (*135*). The large size of the consortium allows for the effective enrichment of *Altiarchaeota* on polyethylene nets that can consist of ~98% of archaeal cells and ~2% bacteria (*196*). Other representatives of the *Altiarchaeota* occur in almost single-member biofilms (~5% bacteria, ~95% *Altiarchaeota*) in sulfidic springs (*192*, *193*).

Notably, *Altiarchaeota* have not only been found in symbiosis with bacteria but represent the hosts of members of the *Huberarchaeota*, a recent addition to the DPANN superphylum (*178*, *197*). Similar to other DPANN archaea, known members of the *Huberarchaeota* have reduced genomes and lack proteins related to energy metabolism, regeneration of redox equivalents and nucleotide biosynthesis indicating that they depend on a variety of compounds from their hosts.

The first insights into the metabolism of the Altiarchaeota came from the metagenome-assembled genome (MAG) of *Candidatus* Altiarchaeum hamiconexum, which was reconstructed from a cold, sulfidic spring in Germany (*179*). Genomic analyses suggested this representative is an anaerobic autotroph, potentially capable of growth on carbon dioxide and possibly acetate, formate, and carbon monoxide (*179*). *Ca.* A. hamiconexum is a biofilm-forming, nonmotile organism with coccoid cells (0.4–0.7 µM in diameter) and a double membrane (*179*). Cells can be surrounded by up to 100 hair-like proteinaceous appendages of 2–3 µM length, so-called hami, which mediate adhesion to various surfaces (*198*). However, representatives of the *Altiarchaeota* from sediments lack genes encoding proteins involved in hami formation and show specific adaptations to their respective environments (*191*).

ARCHAEA AS PART OF THE HUMAN MICROBIOME

For a long time, it was thought that *Archaea* played minor roles in the microbiomes of humans and other mammals and true archaeal pathogens remain to be discovered. The first archaeon associated with humans was described in 1982, the methanogenic *Methanobrevibacter smithii*, which was isolated from human feces (*199*) suggesting that the methane exhaled by a certain proportion of humans may be produced by methanogens residing in the gastrointestinal tract (*200*). Since then, several archaeal species have been identified to be associated with the intestinal, oral, gut, nasal, vaginal, lung, and skin microbiota of both humans and other animals (*201–203*). However, their roles in human health and disease remain poorly understood (*201–205*). In the following, we summarize current knowledge regarding the diversity and function of human-associated archaea.

ORAL ARCHAEOME

Methanogenic archaea are part of the oral archaeome with *Methanobrevibacter oralis* being the most frequently detected species (*205*, *206*). Notably, *M. oralis* seems to be correlated with periodontitis severity, supporting a potential pathogenic role of methanogenic archaea (*206–208*). While no direct experimental evidence has demonstrated the virulence pattern of *M. oralis* and other oral archaeal species, the unique metabolism of methanogenic archaea provides insight into possible drivers of oral disease. For instance, methanogens in periodontal pockets may serve as an H_2 sink, which would favor the proliferation of syntrophic pathogenic microbes (*206–209*). Recent investigation into microbial communities in the oral cavity has shown significant positive correlations between the abundance of methanogens with that of *Prevotella intermedia*, a known bacterial pathogen involved in periodontal infections such as periodontitis, gingivitis, and necrotizing ulcerative gingivitis (*208*).

The relationship between these two groups in periodontal pockets is still unknown, but indirect and direct associations between the methanogens and the local environment may be driving the proliferation of *P. intermedia* through a series of possible syntrophic interactions (*208*). A current key research interest is to further determine the immediate role of archaea in the pathogenesis of periodontal infections (*206*, *210*).

GUT ARCHAEOME

To date, three species of methanogenic archaea have been cultivated and isolated from gut-derived samples, i.e., from human stool: *M. smithii*, *Methanosphaera stadtmanae*, and *Methanomassiliicoccus luminyensis* (*95*, *199*, *211*). With the help of molecular tools, two candidate-species, *Candidatus Methanomassiliicoccus intestinalis* and *Candidatus Methano methylophilus alvus*, in addition to several unknown members of the orders *Methanosarcinales*, *Methanobacteriales*, *Methanococcales*, *Methanomicrobiales*, and *Methanopyrales*, have been shown to inhabit the human gastrointestinal tract (*202*). Further, the presence of methanogens in biopsy samples suggests that they may be associated with the mucosal lining in addition to their presumed presence in the lumen (*202*). *M. smithii* is the major archaeal component in the human gut, while *M. stadtmanae* and *M. luminyensis* are less frequently detected species (*201*, *202*) and appear to play an important role as H_2-consumers in the complex microbial ecosystem of the gut (*201*, *205*, *209*). Fermentative microorganisms produce short-chain fatty acids and H_2, the former being consumed by the host and the latter being scavenged and consumed by the archaea. This removal of H_2 from the system by methanogens makes the fermentative processes kinetically more favorable and continuously drives this cyclical syntrophy (*201*, *202*, *205*). Furthermore, there is evidence that methanogens may be involved in inflammatory bowel disease, irritable bowel syndrome, colorectal cancer, diverticulosis, and obesity (*201*, *205*). However, it is unclear whether methanogens directly or indirectly contribute to the development of gastrointestinal disorders and without doubt, more research is needed to unravel the role of archaea in intestinal disorders (*204*, *212*). For instance, it has also been suggested that some human-associated archaea may be mutualistic, providing health benefits and influencing host metabolism (*202*, *213*).

Not all gut-associated archaea are methanogens however (*202*). For instance, a halophilic archaeon belonging to the halobacteria, *Haloferax massiliensis*, was recently isolated from a human stool sample, reigniting the debate over whether halophiles may colonize the gut (*214*, *215*). Other studies have revealed a diversity of halobacteria-related sequences in fecal samples collected from healthy Korean people, with analyzed sequences representing *Halorubrum alimentarium* and *Halorubrum koreense*. Interestingly, both *H. alimentarium* and *H. koreense* had previously been isolated from salt-fermented sea foods suggesting native cuisine and eating habits may contribute to the propensity of these organisms in the gut environment (*201*, *216*).

GLOBAL HUMAN ARCHAEOME

Technological advancements in high-throughput sequencing have further improved insights into the human microbiome and revealed unexpected diversity of representatives from archaeal phyla that had not previously been detected in human habitats, including members of the DPANN archaea. In particular, members of the *Woesearchaeota* appear to be present in the human lung, and while it is speculated that they may exhibit parasitic/ symbiotic lifestyles, their environmental role remains unclear (*202*). Analytical exploration into the distribution of archaeal signatures in the human body has revealed site-specific patterns that shape a

preliminary biogeographical view of the human archaeome: (1) *Thaumarchaeota* on the skin, (2) methanogenic *Euryarchaeota* in the gut, (3) mixed skin-gut nasal archaeal communities, and (4) *Woesearchaeota* inhabiting the lungs (*202*).

While *M. smithii*, *M. oralis*, *M. stadtmanae*, *M. luminyensis*, and *H. massiliensis* are the only archaeal species that have been successfully isolated and cultivated from human habitats, efforts are underway to culture more archaeal species associated with humans in order to better understand their roles as potential pathogens or commensal members with potentially positive physiological impacts. For instance, a major step toward a better understanding of the function and dynamics of human-associated archaea may be gained through the Human Microbiome Project (*209*, *217*).

Concurrent with efforts to culture archaeal species infecting humans and elucidate their potential roles in human pathogenesis, there are several initiatives aiming to identify antimicrobial agents that are effective against *Archaea*. Research shows that archaea are resistant to antibiotics used to treat bacterial infections, in part due to morphological and physiological features that impede the action of many bacterial-targeting antimicrobial agents. *In vivo* and *in vitro* experiments have indicated susceptibility of several archaeal groups to certain protein synthesis inhibitors, including fusidic acid and imidazole derivatives (*218*). Nonetheless, antibiotic-resistant archaea may become indirectly susceptible to antimicrobial treatments when relying on chemically susceptible bacterial partners within their complex communities. To date, there are a limited number of antimicrobials that target archaea directly. Greater exploration into archaea as causative agents of human disease would also require further investigation into antiarchaeal compounds and treatments (*210*, *218*).

SUMMARY

Thought to be of limited ecological relevance originally, *Archaea* are now known to inhabit a wide range of ecosystems and to play a key role in major biogeochemical cycles (*8*). Furthermore, *Archaea* have proven to be of fundamental importance for our understanding of the evolution of complex eukaryotic cells (*10*) and have emerged as important model systems. Notably, representatives of the *Archaea* are now known to form a stable part of the human microbiome and may even be involved in disease. Unique metabolic and cellular features of *Archaea* are being utilized in a variety of biotechnological applications as well as the development of novel adjuvants in the use of vaccines utilizing the unique membrane lipids of archaeal membranes (*219*). Considering that a large fraction of *Archaea* of high-taxonomic rank likely still awaits discovery (*5*), the coming years will certainly witness further insights into the role of *Archaea* in ecological food webs, the evolution of life and human biology.

FUNDING

This work was supported by a grant of the Swedish Research Council (VR starting grant 2016-03559 to Anja Spang), the NWO-I foundation of the Netherlands Organisation for Scientific Research (WISE fellowship to Anja Spang). Furthermore, Laura Eme is cur- rently supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 803151).

REFERENCES

1. C. R. Woese, G. E. Fox, Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5088–5090 (1977).

2. C. R. Woese, O. Kandler, M. L. Wheelis, Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 4576–4579 (1990).

3. N. Sangwan, F. Xia, J. A. Gilbert, Recovering complete and draft population genomes from metagenome datasets. *Microbiome* **4**, 8 (2016).

4. T. Woyke, D. F. R. Doud, F. Schulz, The trajectory of microbial single-cell sequencing. *Nat. Methods* **14**, 1045–1054 (2017).

5. A. Spang, E. F. Caceres, T. J. G. Ettema, Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. *Science* **357**, eaaf3883 (2017).

6. C. J. Castelle, J. F. Banfield, Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. *Cell* **172**, 1181–1197 (2018).

7. P. S. Adam, G. Borrel, C. Brochier-Armanet, S. Gribaldo, The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. *ISME J.* **11**, 2407–2425 (2017).

8. P. Offre, A. Spang, C. Schleper, Archaea in biogeochemical cycles. *Annu. Rev. Microbiol.* **67**, 437–457 (2013).

9. R. K. Thauer, A.-K. Kaster, H. Seedorf, W. Buckel, R. Hedderich, Methanogenic archaea: ecologically relevant differences in energy conservation. *Nat. Rev. Microbiol.* **6**, 579–591 (2008).

10. L. Eme, A. Spang, J. Lombard, C. W. Stairs, T. J. G. Ettema, Archaea and the origin of eukaryotes. *Nat. Rev. Microbiol.* **16**, 120 (2018).

11. P. López-García, D. Moreira, Open questions on the origin of eukaryotes. *Trends Ecol. Evol.* **30**, 697–708 (2015).

12. W. F. Martin, S. Garg, V. Zimorski, Endosymbiotic theories for eukaryote origin. (2015). https://doi.org/10.1098/rstb.2014.0330.

13. L. Guy, J. H. Saw, T. J. G. Ettema, The archaeal legacy of eukaryotes: a phylogenomic perspective. *Cold Spring Harb. Perspect. Biol.* **6**, a016022 (2014).

14. N. Iwabe, K. Kuma, M. Hasegawa, S. Osawa, T. Miyata, Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 9355–9359 (1989).

15. T. A. Williams, P. G. Foster, C. J. Cox, T. M. Embley, An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* **504**, 231–236 (2013).

16. L. Guy, T. J. G. Ettema, The archaeal "TACK" superphylum and the origin of eukaryotes. *Trends Microbiol.* **19**, 580–587 (2011).

17. A. Spang, J. H. Saw, S. L. Jørgensen, K. Zaremba-Niedzwiedzka, J. Martijn, A. E. Lind, R. van Eijk, C. Schleper, L. Guy, T. J. G. Ettema, Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).

18. K. Zaremba-Niedzwiedzka, E. F. Caceres, J. H. Saw, D. Bäckström, L. Juzokaite, E. Vancaester, K. W. Seitz, K. Anantharaman, P. Starnawski, K. U. Kjeldsen, M. B. Stott, T. Nunoura, J. F. Banfield, A. Schramm, B. J. Baker, A. Spang, T. J. G. Ettema, Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).

19. F. L. Sousa, S. Neukirchen, J. F. Allen, N. Lane, W. F. Martin, Lokiarchaeon is hydrogen dependent. *Nat. Microbiol.* **1**, 16034 (2016).

20. A. Spang, C. W. Stairs, N. Dombrowski, L. Eme, J. Lombard, E. F. Caceres, C. Greening, B. J. Baker, T. J. G. Ettema, Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. *Nat Microbiol* **4**, 1138–1148 (2019).

21. P.-A. Bulzu, A.-Ş. Andrei, M. M. Salcher, M. Mehrshad, K. Inoue, H. Kandori, O. Beja, R. Ghai, H. L. Banciu, Casting light on Asgardarchaeota metabolism in a sunlit microoxic niche. *Nat. Microbiol.* **4**, 1129–1137 (2019).

22. H. Imachi, M. K. Nobu, N. Nakahara, Y. Morono, M. Ogawara, Y. Takaki, Y. Takano, K. Uematsu, T. Ikuta, M. Ito, Y. Matsui, M. Miyazaki, K. Murata, Y. Saito, S. Sakai, C. Song, E. Tasumi, Y. Yamanaka, T. Yamaguchi, Y. Kamagata, H. Tamaki, K. Takai, Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**, 519–525 (2020).

23. T. A. Williams, C. J. Cox, P. G. Foster, G. J. Szöllősi, T. M. Embley, Phylogenomics provides robust support for a two-domains tree of life. *Nat Ecol Evol* **4**, 138–147 (2020).

24. K. S. Makarova, E. V. Koonin, Archaeology of eukaryotic DNA replication. *Cold Spring Harb. Perspect. Biol.* **5**, a012963 (2013).

25. K. Raymann, P. Forterre, C. Brochier-Armanet, S. Gribaldo, Global phylogenomic analysis disentangles the complex evolutionary history of DNA replication in archaea. *Genome Biol. Evol.* **6**, 192–212 (2014).

26. K. S. Makarova, M. Krupovic, E. V. Koonin, Evolution of replicative DNA polymerases in archaea and their contributions to the eukaryotic replication machinery. *Front. Microbiol.* **5**, 354 (2014).

27. F. Mattiroli, S. Bhattacharyya, P. N. Dyer, A. E. White, K. Sandman, B. W. Burkhart, K. R. Byrne, T. Lee, N. G. Ahn, T. J. Santangelo, J. N. Reeve, K. Luger, Structure of histone-based chromatin in Archaea. *Science* **357**, 609–612 (2017).

28. S. Bhattacharyya, F. Mattiroli, K. Luger, Archaeal DNA on the histone merry-go-round. *FEBS J.* **285**, 3168–3174 (2018).

29. F. Werner, D. Grohmann, Evolution of multisubunit RNA polymerases in the three domains of life. *Nat. Rev. Microbiol.* **9**, 85–98 (2011).

30. N. Yutin, K. S. Makarova, Y. I. Wolf, E. V. Koonin, Phylogenomics of prokaryotic ribosomal proteins. *Genome Biology* **12**, P30 (2011).

31. J. A. Krzycki, The direct genetic encoding of pyrrolysine. *Curr. Opin. Microbiol.* **8**, 706–712 (2005).

32. H. Hartman, A. Fedorov, The origin of the eukaryotic cell: a genomic investigation. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 1420–1425 (2002).

33. C. Brochier-Armanet, S. Gribaldo, P. Forterre, A DNA topoisomerase IB in Thaumarchaeota testifies for the presence of this enzyme in the last common ancestor of Archaea and Eucarya. *Biol. Direct* **3**, 54 (2008).

34. E. V. Koonin, N. Yutin, The dispersed archaeal eukaryome and the complex archaeal ancestor of eukaryotes. *Cold Spring Harb. Perspect. Biol.* **6**, a016188 (2014).

35. J. H. Saw, A. Spang, K. Zaremba-Niedzwiedzka, L. Juzokaite, J. A. Dodsworth, S. K. Murugapiran, D. R. Colman, C. Takacs-Vesbach, B. P. Hedlund, L. Guy, T. J. G. Ettema, Exploring microbial dark matter to resolve the deep archaeal ancestry of eukaryotes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140328 (2015).

36. R. Y. Samson, T. Obita, S. M. Freund, R. L. Williams, S. D. Bell, A role for the ESCRT system in cell division in archaea. *Science* **322**, 1710–1713 (2008).

37. A.-C. Lindås, E. A. Karlsson, M. T. Lindgren, T. J. G. Ettema, R. Bernander, A unique cell division machinery in the Archaea. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 18942–18946 (2008).

38. T. J. Ettema, R. Bernander, Cell division and the ESCRT complex: A surprise from the archaea. *Commun. Integr. Biol.* **2**, 86–88 (2009).

39. E. A. Pelve, A.-C. Lindås, W. Martens-Habbena, J. R. de la Torre, D. A. Stahl, R. Bernander, Cdv-based cell division and cell cycle organization in the thaumarchaeon Nitrosopumilus maritimus. *Mol. Microbiol.* **82**, 555–566 (2011).

40. T. J. G. Ettema, A.-C. Lindås, R. Bernander, An actin-based cytoskeleton in archaea. *Mol. Microbiol.* **80**, 1052–1061 (2011).

41. N. Yutin, E. V. Koonin, Archaeal origin of tubulin. *Biol. Direct* **7**, 10 (2012).

42. T. Nunoura, Y. Takaki, J. Kakuta, S. Nishi, J. Sugahara, H. Kazama, G.-J. Chee, M. Hattori, A. Kanai, H. Atomi, K. Takai, H. Takami, Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. *Nucleic Acids Res.* **39**, 3204–3223 (2011).

43. C. M. Klinger, A. Spang, J. B. Dacks, T. J. G. Ettema, Tracing the Archaeal origins of eukaryotic membrane-trafficking system building blocks. *Mol. Biol. Evol.* **33**, 1528–1541 (2016).

44. C. Akıl, R. C. Robinson, Genomes of Asgard archaea encode profilins that regulate actin. *Nature* **562**, 439–443 (2018).

45. L. Villanueva, S. Schouten, J. S. S. Damsté, Phylogenomic analysis of lipid biosynthetic genes of Archaea shed light on the 'lipid divide.' *Environ. Microbiol.* **19**, 54–69 (2017).

46. Y. Koga, H. Morii, Recent advances in structural research on ether lipids from archaea including comparative and physiological aspects. *Biosci. Biotechnol. Biochem.* **69**, 2019–2034 (2005).

47. A. Klingl, S-layer and cytoplasmic membrane - exceptions from the typical archaeal cell wall with a focus on double membranes. *Front. Microbiol.* **5**, 624 (2014).

48. D. H. Parks, M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil, P. Hugenholtz, A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).

49. G. Borrel, P. S. Adam, S. Gribaldo, Methanogenesis and the Wood-Ljungdahl Pathway: An Ancient, Versatile, and Fragile Association. *Genome Biol. Evol.* **8**, 1706–1711 (2016).

50. P. N. Evans, J. A. Boyd, A. O. Leu, B. J. Woodcroft, D. H. Parks, P. Hugenholtz, G. W. Tyson, An evolving view of methane metabolism in the Archaea. *Nat. Rev. Microbiol.* **17**, 219–232 (2019).

51. J. G. Ferry, How to make a living by exhaling methane. *Annu. Rev. Microbiol.* **64**, 453–473 (2010).

52. K. Knittel, A. Boetius, Anaerobic oxidation of methane: progress with an unknown process. *Annu. Rev. Microbiol.* **63**, 311–334 (2009).

53. K. Raymann, C. Brochier-Armanet, S. Gribaldo, The two-domain tree of life is linked to a new root for the Archaea. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 6670–6675 (2015).

54. T. A. Williams, G. J. Szöllősi, A. Spang, P. G. Foster, S. E. Heaps, B. Boussau, T. J. G. Ettema, T. M. Embley, Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E4602–E4611 (2017).

55. J.-L. Garcia, B. Ollivier, W. B. Whitman, "The Order Methanomicrobiales" in *The Prokaryotes* (Springer New York, New York, NY, 2006), pp. 208–230.

56. W. D. Orsi, M. J. L. Coolen, C. Wuchter, L. He, K. D. More, X. Irigoien, G. Chust, C. Johnson, J. D. Hemingway, M. Lee, V. Galy, L. Giosan, Climate oscillations reflected within the microbiome of Arabian Sea sediments. *Sci. Rep.* **7** (2017).

57. M. F. Haroon, S. Hu, Y. Shi, M. Imelfort, J. Keller, P. Hugenholtz, Z. Yuan, G. W. Tyson, Anaerobic oxidation of methane coupled to nitrate reduction in a novel archaeal lineage. *Nature* **500**, 567–570 (2013).

58. K. F. Ettwig, M. K. Butler, D. Le Paslier, E. Pelletier, S. Mangenot, M. M. M. Kuypers, F. Schreiber, B. E. Dutilh, J. Zedelius, D. de Beer, J. Gloerich, H. J. C. T. Wessels, T. van Alen, F. Luesken, M. L. Wu, K. T. van de Pas-Schoonen, H. J. M. Op den Camp, E. M. Janssen-Megens, K.-J. Francoijs, H. Stunnenberg, J. Weissenbach, M. S. M. Jetten, M. Strous, Nitrite-driven anaerobic methane oxidation by oxygenic bacteria. *Nature* **464**, 543–548 (2010).

59. K. F. Ettwig, B. Zhu, D. Speth, J. T. Keltjens, M. S. M. Jetten, B. Kartal, Archaea catalyze iron-dependent anaerobic oxidation of methane. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 12792–12796 (2016).

60. S. E. McGlynn, G. L. Chadwick, C. P. Kempes, V. J. Orphan, Single cell activity reveals direct electron transfer in methanotrophic consortia. *Nature* **526**, 531–535 (2015).

61. W. D. Orsi, Ecology and evolution of seafloor and subseafloor microbial communities. *Nat. Rev. Microbiol.* **16**, 671–683 (2018).

62. P. H. A. Timmers, C. U. Welte, J. J. Koehorst, C. M. Plugge, M. S. M. Jetten, A. J. M. Stams, Reverse methanogenesis and respiration in methanotrophic Archaea. *Archaea* **2017**, 1–22 (2017).

63. K. Knittel, A. Boetius, Anaerobic oxidation of methane: progress with an unknown process (b). *Annu. Rev. Microbiol.* **63**, 311–334 (2009).

64. G. Wegener, V. Krukenberg, D. Riedel, H. E. Tegetmeyer, A. Boetius, Intercellular wiring enables electron transfer between methanotrophic archaea and bacteria. *Nature* **526**, 587–590 (2015).

65. S. Sakai, H. Imachi, S. Hanada, A. Ohashi, H. Harada, Y. Kamagata, Methanocella paludicola gen. nov., sp. nov., a methane-producing archaeon, the first isolate of the lineage "Rice Cluster I", and proposal of the new archaeal order Methanocellales ord. nov. *Int. J. Syst. Evol. Microbiol.* **58**, 929–936 (2008).

66. Y. Lu, R. Conrad, In situ stable isotope probing of methanogenic archaea in the rice rhizosphere. *Science* **309**, 1088–1090 (2005).

67. S. Sakai, H. Imachi, Y. Sekiguchi, A. Ohashi, H. Harada, Y. Kamagata, Isolation of key methanogens for global methane emission from rice paddy fields: a novel isolate affiliated with the clone cluster rice cluster I. *Appl. Environ. Microbiol.* **73**, 4326–4331 (2007).

68. S. Sakai, R. Conrad, W. Liesack, H. Imachi, Methanocella arvoryzae sp. nov., a hydrogenotrophic methanogen isolated from rice field soil. *Int. J. Syst. Evol. Microbiol.* **60**, 2918–2923 (2010).

69. Z. Lü, Y. Lu, Methanocella conradii sp. nov., a thermophilic, obligate hydrogenotrophic methanogen, isolated from Chinese rice field soil. *PLoS One* **7**, e35279 (2012).

70. B. N. Orcutt, S. B. Joye, S. Kleindienst, K. Knittel, A. Ramette, A. Reitz, V. Samarkin, T. Treude, A. Boetius, Impact of natural oil and higher hydrocarbons on microbial diversity, distribution, and activity in Gulf of Mexico cold-seep sediments. *Deep Sea Res. Part 2 Top. Stud. Oceanogr.* **57**, 2008–2021 (2010).

71. R. Laso-Pérez, G. Wegener, K. Knittel, F. Widdel, K. J. Harding, V. Krukenberg, D. V. Meier, M. Richter, H. E. Tegetmeyer, D. Riedel, H.-H. Richnow, L. Adrian, T. Reemtsma, O. J. Lechtenfeld, F. Musat, Thermophilic archaea activate butane via alkyl-coenzyme M formation. *Nature* **539**, 396–401 (2016).

72. R. Laso-Pérez, V. Krukenberg, F. Musat, G. Wegener, Establishing anaerobic hydrocarbon-degrading enrichment cultures of microorganisms under strictly anoxic conditions. *Nat. Protoc.* **13**, 1310–1330 (2018).

73. R. Laso-Pérez, C. Hahn, D. M. van Vliet, H. E. Tegetmeyer, F. Schubotz, N. T. Smit, T. Pape, H. Sahling, G. Bohrmann, A. Boetius, K. Knittel, G. Wegener, Anaerobic degradation of non-methane alkanes by "Candidatus Methanoliparia" in hydrocarbon seeps of the Gulf of Mexico. *MBio* **10** (2019).

74. D. V. Fedosov, D. A. Podkopaeva, M. L. Miroshnichenko, E. A. Bonch-Osmolovskaya, A. V. Lebedinsky, M. Y. Grabovich, Investigation of the catabolism of acetate and peptides in the new anaerobic thermophilic bacterium Caldithrix abyssi. *Microbiology* **75**, 119–124 (2006).

75. D. R. Boone, R. W. Castenholz, *Bergey's Manual of Systematic Bacteriology* (Springer-Verlag, New York, ed. 2, 2001)vol. 1.

76. J. A. Boyd, S. P. Jungbluth, A. O. Leu, P. N. Evans, B. J. Woodcroft, G. L. Chadwick, V. J. Orphan, J. P. Amend, M. S. Rappé, G. W. Tyson, Divergent methyl-coenzyme M reductase genes in a deep-subseafloor Archaeoglobi. *ISME J.* **13**, 1269–1279 (2019).

77. Y. Wang, X. Feng, V. P. Natarajan, X. Xiao, F. Wang, Diverse anaerobic methane- and multi-carbon alkane-metabolizing archaea coexist and show activity in Guaymas Basin hydrothermal sediment. *Environ. Microbiol.* **21**, 1344–1355 (2019).

78. G. Borrel, P. S. Adam, L. J. McKay, L.-X. Chen, I. N. Sierra-García, C. M. K. Sieber, Q. Letourneur, A. Ghozlane, G. L. Andersen, W.-J. Li, S. J. Hallam, G. Muyzer, V. M. de Oliveira, W. P. Inskeep, J. F. Banfield, S. Gribaldo, Wide diversity of methane and short-chain alkane metabolisms in uncultured archaea. *Nat. Microbiol.* **4**, 603–613 (2019).

79. J. Soppa, From replication to cultivation: hot news from Haloarchaea. *Curr. Opin. Microbiol.* **8**, 737–744 (2005).

80. A. E. Walsby, Archaea with square cells. *Trends Microbiol.* **13**, 193–195 (2005).

81. D. Y. Sorokin, A. Y. Merkel, B. Abbas, K. S. Makarova, W. I. C. Rijpstra, M. Koenen, J. S. Sinninghe Damsté, E. A. Galinski, E. V. Koonin, M. C. M. van Loosdrecht, Methanonatronarchaeum thermophilum gen. nov., sp. nov. and "Candidatus Methanohalarchaeum thermophilum", extremely halo(natrono) philic methyl-reducing methanogens from hypersaline lakes comprising a new euryarchaeal class Methanonatronarchaeia classis nov. *Int. J. Syst. Evol. Microbiol.* **68**, 2199–2208 (2018).

82. D. Y. Sorokin, K. S. Makarova, B. Abbas, M. Ferrer, P. N. Golyshin, E. A. Galinski, S. Ciordia, M. C. Mena, A. Y. Merkel, Y. I. Wolf, M. C. M. van Loosdrecht, E. V. Koonin, Discovery of extremely halophilic, methyl-reducing euryarchaea provides insights into the evolutionary origin of methanogenesis. *Nat Microbiol* **2**, 17081 (2017).

83. W. Eder, M. Schmidt, M. Koch, D. Garbe-Schönberg, R. Huber, Prokaryotic phylogenetic diversity and corresponding geochemical data of the brine-seawater interface of the Shaban Deep, Red Sea. *Environ. Microbiol.* **4**, 758–763 (2002).

84. M. Aouad, G. Borrel, C. Brochier-Armanet, S. Gribaldo, Evolutionary placement of Methanonatronarchaeia, *Nature microbiology*. **4** (2019)pp. 558–559.

85. J. A. Fuhrman, K. McCallum, A. A. Davis, Novel major archaebacterial group from marine plankton. *Nature* **356**, 148–149 (1992).

86. E. F. DeLong, Archaea in coastal marine environments. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 5685–5689 (1992).

87. G. Borrel, N. Parisot, H. M. B. Harris, E. Peyretaillade, N. Gaci, W. Tottey, O. Bardot, K. Raymann, S. Gribaldo, P. Peyret, P. W. O'Toole, J.-F. Brugère, Comparative genomics highlights the unique biology of Methanomassiliicoccales, a Thermoplasmatales-related seventh order of methanogenic archaea that encodes pyrrolysine. *BMC Genomics* **15**, 679 (2014).

88. O. V. Golyshina, I. V. Kublanov, H. Tran, A. A. Korzhenkov, H. Lünsdorf, T. Y. Nechitaylo, S. N. Gavrilov, S. V. Toshchakov, P. N. Golyshin, Biology of archaea from a novel family Cuniculiplasmataceae (Thermoplasmata) ubiquitous in hyperacidic environments. *Sci. Rep.* **6** (2016).

89. C. Schleper, G. Puehler, I. Holz, A. Gambacorta, D. Janekovic, U. Santarius, H. P. Klenk, W. Zillig, Picrophilus gen. nov., fam. nov.: a novel aerobic, heterotrophic, thermoacidophilic genus and family comprising archaea capable of growth around pH 0. *J. Bacteriol.* **177**, 7050–7059 (1995).

90. R. Edwards, D. P. Dixon, V. Walbot, Plant glutathione S-transferases: enzymes with multiple functions in sickness and in health. *Trends Plant Sci.* **5**, 193–198 (2000).

91. O. V. Golyshina, Environmental, biogeographic, and biochemical patterns of archaea of the family Ferroplasmaceae. *Appl. Environ. Microbiol.* **77**, 5071–5078 (2011).

92. A.-L. Reysenbach, Y. Liu, A. B. Banta, T. J. Beveridge, J. D. Kirshtein, S. Schouten, M. K. Tivey, K. L. Von Damm, M. A. Voytek, A ubiquitous thermoacidophilic archaeon from deep-sea hydrothermal vents. *Nature* **442**, 444–447 (2006).

93. G. E. Flores, I. D. Wagner, Y. Liu, A.-L. Reysenbach, Distribution, abundance, and diversity patterns of the thermoacidophilic "deep-sea hydrothermal vent euryarchaeota 2." *Front. Microbiol.* **3**, 47 (2012).

94. K. Takai, K. Horikoshi, Genetic diversity of archaea in deep-sea hydrothermal vent environments. *Genetics* **152**, 1285–1297 (1999).

95. B. Dridi, M.-L. Fardeau, B. Ollivier, D. Raoult, M. Drancourt, Methanomassiliicoccus luminyensis gen. nov., sp. nov., a methanogenic archaeon isolated from human faeces. *Int. J. Syst. Evol. Microbiol.* **62**, 1902–1907 (2012).

96. L. Kröninger, J. Gottschling, U. Deppenmeier, Growth characteristics of Methanomassiliicoccus luminyensis and expression of methyltransferase encoding genes. *Archaea* **2017**, 2756573 (2017).

97. G. Borrel, H. M. B. Harris, W. Tottey, A. Mihajlovski, N. Parisot, E. Peyretaillade, P. Peyret, S. Gribaldo, P. W. O'Toole, J.-F. Brugère, Genome sequence of "Candidatus Methanomethylophilus alvus" Mx1201, a methanogenic archaeon from the human gut belonging to a seventh order of methanogens. *J. Bacteriol.* **194**, 6944–6945 (2012).

98. K. Lang, J. Schuldes, A. Klingl, A. Poehlein, R. Daniel, A. Brunea, New mode of energy metabolism in the seventh order of methanogens as revealed by comparative genome analysis of "Candidatus methanoplasma termitum." *Appl. Environ. Microbiol.* **81**, 1338–1352 (2015).

99. M. Poulsen, C. Schwab, B. B. Jensen, R. M. Engberg, A. Spang, N. Canibe, O. Højberg, G. Milinovich, L. Fragner, C. Schleper, W. Weckwerth, P. Lund, A. Schramm, T. Urich, Methylotrophic methanogenic Thermoplasmata implicated in reduced methane emissions from bovine rumen. *Nat. Commun.* **4**, 1428 (2013).

100. G. Borrel, H. M. B. Harris, N. Parisot, N. Gaci, W. Tottey, A. Mihajlovski, J. Deane, S. Gribaldo, O. Bardot, E. Peyretaillade, P. Peyret, P. W. O'Toole, J.-F. Brugère, Genome sequence of "Candidatus Methanomassiliicoccus intestinalis" Issoire-Mx1, a third Thermoplasmatales-related methanogenic archaeon from human feces. *Genome Announc.* **1** (2013).

101. C. Rinke, F. Rubino, L. F. Messer, N. Youssef, D. H. Parks, M. Chuvochina, M. Brown, T. Jeffries, G. W. Tyson, J. R. Seymour, P. Hugenholtz, A phylogenomic and ecological analysis of the globally abundant Marine Group II archaea (*Ca.* Poseidoniales ord. nov.). *ISME J.* **13**, 663–675 (2019).

102. J. A. Fuhrman, A. A. Davis, Widespread Archaea and novel Bacteria from the deep sea as shown by 16S rRNA gene sequences. *Mar. Ecol. Prog. Ser.* **150**, 275–285 (1997).

103. R. Massana, A. E. Murray, C. M. Preston, E. F. DeLong, Vertical distribution and phylogenetic characterization of marine planktonic Archaea in the Santa Barbara Channel. *Appl. Environ. Microbiol.* **63**, 50–56 (1997).

104. E. Teira, T. Reinthaler, A. Pernthaler, J. Pernthaler, G. J. Herndl, Combining catalyzed reporter deposition-fluorescence in situ hybridization and microautoradiography to detect substrate utilization by bacteria and Archaea in the deep ocean. *Appl. Environ. Microbiol.* **70**, 4411–4414 (2004).

105. L. H. Orellana, T. Ben Francis, K. Krüger, H. Teeling, M.-C. Müller, B. M. Fuchs, K. T. Konstantinidis, R. I. Amann, Niche differentiation among annually recurrent coastal Marine Group II Euryarchaeota. *ISME J.* **13**, 3024–3036 (2019).

106. N.-U. Frigaard, A. Martinez, T. J. Mincer, E. F. DeLong, Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature* **439**, 847–850 (2006).

107. V. Iverson, R. M. Morris, C. D. Frazar, C. T. Berthiaume, R. L. Morales, E. V. Armbrust, Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science* **335**, 587–590 (2012).

108. P. Deschamps, Y. Zivanovic, D. Moreira, F. Rodriguez-Valera, P. López-García, Pangenome evidence for extensive interdomain horizontal transfer affecting lineage core and shell genes in uncultured planktonic thaumarchaeota and euryarchaeota. *Genome Biol. Evol.* **6**, 1549–1563 (2014).

109. M. Li, B. J. Baker, K. Anantharaman, S. Jain, J. A. Breier, G. J. Dick, Genomic and transcriptomic evidence for scavenging of diverse organic compounds by widespread deep-sea archaea. *Nat. Commun.* **6**, 8933 (2015).

110. B. J. Tully, Metabolic diversity within the globally abundant Marine Group II Euryarchaea offers insight into ecological patterns. *Nat. Commun.* **10**, 271 (2019).

111. J. M. Haro-Moreno, F. Rodriguez-Valera, P. López-García, D. Moreira, A.-B. Martin-Cuadrado, New insights into marine group III Euryarchaeota, from dark to light. *ISME J.* **11**, 1102–1117 (2017).

112. M. Li, B. J. Baker, K. Anantharaman, S. Jain, J. A. Breier, G. J. Dick, Genomic and transcriptomic evidence for scavenging of diverse organic compounds by widespread deep-sea archaea (b). *Nat. Commun.* **6**, 8933 (2015).

113. Y. Yokooji, T. Sato, S. Fujiwara, T. Imanaka, H. Atomi, Genetic examination of initial amino acid oxidation and glutamate catabolism in the hyperthermophilic archaeon Thermococcus kodakarensis. *J. Bacteriol.* **195**, 1940–1948 (2013).

114. F. L. Poole 2nd, B. A. Gerwe, R. C. Hopkins, G. J. Schut, M. V. Weinberg, F. E. Jenney Jr, M. W. W. Adams, Defining genes in the genome of the hyperthermophilic archaeon Pyrococcus furiosus: implications for all microbial genomes. *J. Bacteriol.* **187**, 7325–7332 (2005).

115. R. Chouari, D. Le Paslier, P. Daegelen, P. Ginestet, J. Weissenbach, A. Sghir, Novel predominant archaeal and bacterial groups revealed by molecular analysis of an anaerobic sludge digester. *Environ. Microbiol.* **7**, 1104–1115 (2005).

116. M. A. Dojka, P. Hugenholtz, S. K. Haack, N. R. Pace, Microbial diversity in a hydrocarbon- and chlorinated-solvent-contaminated aquifer undergoing intrinsic bioremediation. *Appl. Environ. Microbiol.* **64**, 3869–3877 (1998).

117. M. K. Nobu, T. Narihiro, K. Kuroda, R. Mei, W.-T. Liu, Chasing the elusive Euryarchaeota class WSA2: genomes reveal a uniquely fastidious methyl-reducing methanogen. *ISME J.* **10**, 2478–2487 (2016).

118. J. W. Sahl, M. O. Gary, J. K. Harris, J. R. Spear, A comparative molecular analysis of water-filled limestone sinkholes in north-eastern Mexico. *Environ. Microbiol.* **13**, 226–240 (2011).

119. C. S. Lazar, B. J. Baker, K. W. Seitz, A. P. Teske, Genomic reconstruction of multiple lineages of uncultured benthic archaea suggests distinct biogeochemical roles and ecological niches. *ISME J.* **11**, 1118–1129 (2017).

120. K. Takai, D. P. Moser, M. DeFlaun, T. C. Onstott, J. K. Fredrickson, Archaeal diversity in waters from deep South African gold mines. *Appl. Environ. Microbiol.* **67**, 5750–5760 (2001).

121. R. J. Parkes, G. Webster, B. A. Cragg, A. J. Weightman, C. J. Newberry, T. G. Ferdelman, J. Kallmeyer, B. B. Jørgensen, I. W. Aiello, J. C. Fry, Deep sub-seafloor prokaryotes stimulated at interfaces over geological time. *Nature* **436**, 390–394 (2005).

122. J. F. Biddle, J. S. Lipp, M. A. Lever, K. G. Lloyd, K. B. Sørensen, R. Anderson, H. F. Fredricks, M. Elvert, T. J. Kelly, D. P. Schrag, M. L. Sogin, J. E. Brenchley, A. Teske, C. H. House, K.-U. Hinrichs, Heterotrophic Archaea dominate sedimentary subsurface ecosystems off Peru. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 3846–3851 (2006).

123. B. J. Baker, J. H. Saw, A. E. Lind, C. S. Lazar, K.-U. Hinrichs, A. P. Teske, T. J. G. Ettema, Genomic inference of the metabolism of cosmopolitan subsurface Archaea, Hadesarchaea. *Nat. Microbiol.* **1**, 16002 (2016).

124. Z.-S. Hua, Y.-L. Wang, P. N. Evans, Y.-N. Qu, K. M. Goh, Y.-Z. Rao, Y.-L. Qi, Y.-X. Li, M.-J. Huang, J.-Y. Jiao, Y.-T. Chen, Y.-P. Mao, W.-S. Shu, W. Hozzein, B. P. Hedlund, G. W. Tyson, T. Zhang, W.-J. Li, Insights into the ecological roles and evolution of methyl-coenzyme M reductase-containing hot spring Archaea. *Nat. Commun.* **10**, 4574 (2019).

125. P. W. J. J. van der Wielen, H. Bolhuis, S. Borin, D. Daffonchio, C. Corselli, L. Giuliano, G. D'Auria, G. J. de Lange, A. Huebner, S. P. Varnavas, J. Thomson, C. Tamburini, D. Marty, T. J. McGenity, K. N. Timmis, BioDeep Scientific Party, The enigma of prokaryotic life in deep hypersaline anoxic basins. *Science* **307**, 121–123 (2005).

126. Y. Guan, T. Hikmawan, A. Antunes, D. Ngugi, U. Stingl, Diversity of methanogens and sulfate-reducing bacteria in the interfaces of five deep-sea anoxic brines of the Red Sea. *Res. Microbiol.* **166**, 688–699 (2015).

127. R. Mwirichia, I. Alam, M. Rashid, M. Vinu, W. Ba-Alawi, A. Anthony Kamau, D. Kamanda Ngugi, M. Göker, H.-P. Klenk, V. Bajic, U. Stingl, Metabolic traits of an uncultured archaeal lineage -MSBL1- from brine pools of the Red Sea. *Sci. Rep.* **6** (2016).

128. C. Vetriani, H. W. Jannasch, B. J. MacGregor, D. A. Stahl, A. L. Reysenbach, Population structure and phylogenetic characterization of marine benthic Archaea in deep-sea sediments. *Appl. Environ. Microbiol.* **65**, 4375–4384 (1999).

129. S. P. Jungbluth, J. P. Amend, M. S. Rappé, Metagenome sequencing and 98 microbial genomes from Juan de Fuca Ridge flank subsurface fluids. *Sci. Data* **4**, 170037 (2017).

130. N. Dombrowski, A. P. Teske, B. J. Baker, Expansive microbial metabolic versatility and biodiversity in dynamic Guaymas Basin hydrothermal sediments. *Nat. Commun.* **9**, 4999 (2018).

131. Z. Zhou, Y. Liu, W. Xu, J. Pan, Z.-H. Luo, M. Li, Genome- and community-level interaction insights into carbon utilization and element cycling functions of Hydrothermarchaeota in hydrothermal sediment. *mSystems* **5** (2020).

132. S. A. Carr, S. P. Jungbluth, E. A. Eloe-Fadrosh, R. Stepanauskas, T. Woyke, M. S. Rappé, B. N. Orcutt, Carboxydotrophy potential of uncultivated Hydrothermarchaeota from the subseafloor crustal biosphere. *ISME J.* **13**, 1457–1468 (2019).

133. S. Kato, S. Nakano, M. Kouduka, M. Hirai, K. Suzuki, T. Itoh, M. Ohkuma, Y. Suzuki, Metabolic potential of as-yet-uncultured Archaeal lineages of Candidatus Hydrothermarchaeota thriving in deep-sea metal sulfide deposits. *Microbes Environ.* **34**, 293–303 (2019).

134. C. Petitjean, P. Deschamps, P. López-García, D. Moreira, Rooting the domain archaea by phylogenomic analysis supports the foundation of the new kingdom Proteoarchaeota. *Genome Biol. Evol.* **7**, 191–204 (2014).

135. C. Rudolph, G. Wanner, R. Huber, Natural communities of novel archaea and bacteria growing in cold sulfurous springs with a string-of-pearls-like morphology. *Appl. Environ. Microbiol.* **67**, 2336–2344 (2001).

136. H. Huber, M. J. Hohn, R. Rachel, T. Fuchs, V. C. Wimmer, K. O. Stetter, A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* **417**, 63–67 (2002).

137. M. Podar, K. S. Makarova, D. E. Graham, Y. I. Wolf, E. V. Koonin, A.-L. Reysenbach, Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone National Park. *Biol. Direct* **8**, 9 (2013).

138. L. Wurch, R. J. Giannone, B. S. Belisle, C. Swift, S. Utturkar, R. L. Hettich, A.-L. Reysenbach, M. Podar, Genomics-informed isolation and characterization of a symbiotic Nanoarchaeota system from a terrestrial geothermal environment. *Nat. Commun.* **7**, 12115 (2016).

139. J. K. Jarett, S. Nayfach, M. Podar, W. Inskeep, N. N. Ivanova, J. Munson-McGee, F. Schulz, M. Young, Z. J. Jay, J. P. Beam, N. C. Kyrpides, R. R. Malmstrom, R. Stepanauskas, T. Woyke, Single-cell genomics of co-sorted Nanoarchaeota suggests novel putative host associations and diversification of proteins involved in symbiosis. *Microbiome* **6**, 161 (2018).

140. E. St John, G. E. Flores, J. Meneghin, A.-L. Reysenbach, Deep-sea hydrothermal vent metagenome-assembled genomes provide insight into the phylum Nanoarchaeota. *Environ. Microbiol. Rep.* **11**, 262–270 (2019).

141. H. Huber, U. Küper, S. Daxer, R. Rachel, The unusual cell biology of the hyperthermophilic Crenarchaeon Ignicoccus hospitalis. *Antonie Van Leeuwenhoek* **102**, 203–219 (2012).

142. T. Heimerl, J. Flechsler, C. Pickl, V. Heinz, B. Salecker, J. Zweck, G. Wanner, S. Geimer, R. Y. Samson, S. D. Bell, H. Huber, R. Wirth, L. Wurch, M. Podar, R. Rachel, A Complex Endomembrane System in the Archaeon Ignicoccus hospitalis Tapped by Nanoarchaeum equitans. *Front. Microbiol.* **8**, 1072 (2017).

143. M. Könneke, A. E. Bernhard, J. R. de la Torre, C. B. Walker, J. B. Waterbury, D. A. Stahl, Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* **437**, 543–546 (2005).

144. C. Brochier-Armanet, B. Boussau, S. Gribaldo, P. Forterre, Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nat. Rev. Microbiol.* **6**, 245–252 (2008).

145. A. Spang, R. Hatzenpichler, C. Brochier-Armanet, T. Rattei, P. Tischler, E. Spieck, W. Streit, D. A. Stahl, M. Wagner, C. Schleper, Distinct gene set in two different lineages of ammonia-oxidizing archaea supports the phylum Thaumarchaeota. *Trends Microbiol.* **18**, 331–340 (2010).

146. D. A. Stahl, J. R. de la Torre, Physiology and diversity of ammonia-oxidizing archaea. *Annu. Rev. Microbiol.* **66**, 83–101 (2012).

147. J. P. Beam, Z. J. Jay, M. A. Kozubal, W. P. Inskeep, Niche specialization of novel Thaumarchaeota to oxic and hypoxic acidic geothermal springs of Yellowstone National Park. *ISME J.* **8**, 938–951 (2014).

148. S. Kato, T. Itoh, M. Yuki, M. Nagamori, M. Ohnishi, K. Uematsu, K. Suzuki, T. Takashina, M. Ohkuma, Isolation and characterization of a thermophilic sulfur- and iron-reducing thaumarchaeote from a terrestrial acidic hot spring. *ISME J.* **13**, 2465–2474 (2019).

149. C. Rinke, P. Schwientek, A. Sczyrba, N. N. Ivanova, I. J. Anderson, J.-F. Cheng, A. Darling, S. Malfatti, B. K. Swan, E. A. Gies, J. A. Dodsworth, B. P. Hedlund, G. Tsiamis, S. M. Sievert, W.-T. Liu, J. A. Eisen, S. J. Hallam, N. C. Kyrpides, R. Stepanauskas, E. M. Rubin, P. Hugenholtz, T. Woyke, Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).

150. B. P. Hedlund, S. K. Murugapiran, T. W. Alba, A. Levy, J. A. Dodsworth, G. B. Goertz, N. Ivanova, T. Woyke, Uncultivated thermophiles: current status and spotlight on "Aigarchaeota." *Curr. Opin. Microbiol.* **25**, 136–145 (2015).

151. J. P. Beam, Z. J. Jay, M. C. Schmid, D. B. Rusch, M. F. Romine, R. de M. Jennings, M. A. Kozubal, S. G. Tringe, M. Wagner, W. P. Inskeep, Ecophysiology of an uncultivated lineage of Aigarchaeota from an oxic, hot spring filamentous "streamer" community. *ISME J.* **10**, 210–224 (2016).

152. Z.-S. Hua, Y.-N. Qu, Q. Zhu, E.-M. Zhou, Y.-L. Qi, Y.-R. Yin, Y.-Z. Rao, Y. Tian, Y.-X. Li, L. Liu, C. J. Castelle, B. P. Hedlund, W.-S. Shu, R. Knight, W.-J. Li, Genomic inference of the metabolism and evolution of the archaeal phylum Aigarchaeota. *Nat. Commun.* **9**, 2832 (2018).

153. S. M. Barns, C. F. Delwiche, J. D. Palmer, N. R. Pace, Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 9188–9193 (1996).

154. J. G. Elkins, M. Podar, D. E. Graham, K. S. Makarova, Y. Wolf, L. Randau, B. P. Hedlund, C. Brochier-Armanet, V. Kunin, I. Anderson, A. Lapidus, E. Goltsman, K. Barry, E. V. Koonin, P. Hugenholtz, N. Kyrpides, G. Wanner, P. Richardson, M. Keller, K. O. Stetter, A korarchaeal genome reveals insights into the evolution of the Archaea. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 8102–8107 (2008).

155. L. J. McKay, M. Dlakić, M. W. Fields, T. O. Delmont, A. M. Eren, Z. J. Jay, K. B. Klingelsmith, D. B. Rusch, W. P. Inskeep, Co-occurring genomic capacity for anaerobic methane and dissimilatory sulfur metabolisms discovered in the Korarchaeota. *Nat. Microbiol.* **4**, 614–622 (2019).

156. F. Inagaki, M. Suzuki, K. Takai, H. Oida, T. Sakamoto, K. Aoki, K. H. Nealson, K. Horikoshi, Microbial communities associated with geological horizons in coastal subseafloor sediments from the sea of okhotsk. *Appl. Environ. Microbiol.* **69**, 7224–7235 (2003).

157. Z. Zhou, J. Pan, F. Wang, J.-D. Gu, M. Li, Bathyarchaeota: globally distributed metabolic generalists in anoxic environments. *FEMS Microbiol. Rev.* **42**, 639–655 (2018).

158. A. P. Teske, Microbial communities of deep marine subsurface sediments: Molecular and cultivation surveys. *Geomicrobiol. J.* **23**, 357–368 (2006).

159. K. Kubo, K. G. Lloyd, J. F Biddle, R. Amann, A. Teske, K. Knittel, Archaea of the Miscellaneous Crenarchaeotal Group are abundant, diverse and widespread in marine sediments. *ISME J.* **6**, 1949–1965 (2012).

160. T. Yu, W. Wu, W. Liang, M. A. Lever, K.-U. Hinrichs, F. Wang, Growth of sedimentary Bathyarchaeota on lignin as an energy source. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 6022–6027 (2018).

161. Y. He, M. Li, V. Perumal, X. Feng, J. Fang, J. Xie, S. M. Sievert, F. Wang, Genomic and enzymatic evidence for acetogenesis among multiple lineages of the archaeal phylum Bathyarchaeota widespread in marine sediments. *Nat. Microbiol.* **1**, 16035 (2016).

162. P. N. Evans, D. H. Parks, G. L. Chadwick, S. J. Robbins, V. J. Orphan, S. D. Golding, G. W. Tyson, Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science* **350**, 434–438 (2015).

163. M. A. Kozubal, R. E. Macur, Z. J. Jay, J. P. Beam, S. A. Malfatti, S. G. Tringe, B. D. Kocar, T. Borch, W. P. Inskeep, Microbial iron cycling in acidic geothermal springs of yellowstone national park: integrating molecular surveys, geochemical processes, and isolation of novel fe-active microorganisms. *Front. Microbiol.* **3**, 109 (2012).

164. M. A. Kozubal, M. Romine, R. D. Jennings, Z. J. Jay, S. G. Tringe, D. B. Rusch, J. P. Beam, L. A. McCue, W. P. Inskeep, Geoarchaeota: a new candidate phylum in the Archaea from high-temperature acidic iron mats in Yellowstone National Park. *ISME J.* **7**, 622–634 (2013).

165. J. P. Beam, H. C. Bernstein, Z. J. Jay, M. A. Kozubal, R. D. Jennings, S. G. Tringe, W. P. Inskeep, Assembly and succession of iron oxide microbial mat communities in acidic geothermal springs. *Front. Microbiol.* **7**, 25 (2016).

166. L. Guy, A. Spang, J. H. Saw, T. J. G. Ettema, 'Geoarchaeote NAG1' is a deeply rooting lineage of the archaeal order Thermoproteales rather than a new phylum. *ISME J.* **8**, 1353–1357 (2014).

167. I. Vanwonterghem, P. N. Evans, D. H. Parks, P. D. Jensen, B. J. Woodcroft, P. Hugenholtz, G. W. Tyson, Methylotrophic methanogenesis discovered in the archaeal phylum Verstraetearchaeota. *Nat. Microbiol.* **1**, 16170 (2016).

168. B. A. Berghuis, F. B. Yu, F. Schulz, P. C. Blainey, T. Woyke, S. R. Quake, Hydrogenotrophic methanogenesis in archaeal phylum Verstraetearchaeota reveals the shared ancestry of all methanogens. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 5037–5044 (2019).

169. V. V. Kadnikov, A. V. Mardanov, A. V. Beletsky, Y. A. Frank, O. V. Karnachuk, N. V. Ravin, Genome of a member of the candidate Archaeal phylum verstraetearchaeota from a subsurface thermal aquifer revealed pathways of methyl-reducing methanogenesis and fermentative metabolism. *Microbiology* **88**, 316–323 (2019).

170. Z. J. Jay, J. P. Beam, M. Dlakić, D. B. Rusch, M. A. Kozubal, W. P. Inskeep, Marsarchaeota are an aerobic archaeal lineage abundant in geothermal iron oxide microbial mats. *Nat. Microbiol.* **3**, 732–740 (2018).

171. S. L. Jorgensen, B. Hannisdal, A. Lanzén, T. Baumberger, K. Flesland, R. Fonseca, L. Ovreås, I. H. Steen, I. H. Thorseth, R. B. Pedersen, C. Schleper, Correlating microbial community profiles with geochemical data in highly stratified sediments from the Arctic Mid-Ocean Ridge. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E2846-55 (2012).

172. A. Wilke, J. Bischof, T. Harrison, T. Brettin, M. D'Souza, W. Gerlach, H. Matthews, T. Paczian, J. Wilkening, E. M. Glass, N. Desai, F. Meyer, A RESTful API for accessing microbial community data for MG-RAST. *PLoS Comput. Biol.* **11**, e1004008 (2015).

173. K. W. Seitz, C. S. Lazar, K.-U. Hinrichs, A. P. Teske, B. J. Baker, Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. *ISME J.* **10**, 1696–1705 (2016).

174. L. Manoharan, J. A. Kozlowski, R. W. Murdoch, F. E. Löffler, F. L. Sousa, C. Schleper, Metagenomes from coastal marine sediments give insights into the ecological role and cellular features of Loki- and Thorarchaeota. *MBio* **10** (2019).

175. K. W. Seitz, N. Dombrowski, L. Eme, A. Spang, J. Lombard, J. R. Sieber, A. P. Teske, T. J. G. Ettema, B. J. Baker, Asgard archaea capable of anaerobic hydrocarbon cycling. *Nat. Commun.* **10**, 1822 (2019).

176. C. J. Castelle, K. C. Wrighton, B. C. Thomas, L. A. Hug, C. T. Brown, M. J. Wilkins, K. R. Frischkorn, S. G. Tringe, A. Singh, L. M. Markillie, R. C. Taylor, K. H. Williams, J. F. Banfield, Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr. Biol.* **25**, 690–701 (2015).

177. C. J. Castelle, C. T. Brown, K. Anantharaman, A. J. Probst, R. H. Huang, J. F. Banfield, Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat. Rev. Microbiol.* **16**, 629–645 (2018).

178. A. J. Probst, B. Ladd, J. K. Jarett, D. E. Geller-McGrath, C. M. K. Sieber, J. B. Emerson, K. Anantharaman, B. C. Thomas, R. R. Malmstrom, M. Stieglmeier, A. Klingl, T. Woyke, M. C. Ryan, J. F. Banfield, Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface. *Nat. Microbiol.* **3**, 328–336 (2018).

179. A. J. Probst, T. Weinmaier, K. Raymann, A. Perras, J. B. Emerson, T. Rattei, G. Wanner, A. Klingl, I. A. Berg, M. Yoshinaga, B. Viehweger, K.-U. Hinrichs, B. C. Thomas, S. Meck, A. K. Auerbach, M. Heise, A. Schintlmeister, M. Schmid, M. Wagner, S. Gribaldo, J. F. Banfield, C. Moissl-Eichinger, Biology of a widespread uncultivated archaeon that contributes to carbon fixation in the subsurface. *Nat. Commun.* **5**, 5497 (2014).

180. N. Dombrowski, J.-H. Lee, T. A. Williams, P. Offre, A. Spang, Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS Microbiol. Lett.* **366** (2019).

181. C. Brochier, S. Gribaldo, Y. Zivanovic, F. Confalonieri, P. Forterre, Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales? *Genome Biol.* **6**, R42 (2005).

182. U. Jahn, R. Summons, H. Sturt, E. Grosjean, H. Huber, Composition of the lipids of Nanoarchaeum equitans and their origin from its host Ignicoccus sp. strain KIN4/I. *Arch. Microbiol.* **182**, 404–413 (2004).

183. U. Jahn, M. Gallenberger, W. Paper, B. Junglas, W. Eisenreich, K. O. Stetter, R. Rachel, H. Huber, Nanoarchaeum equitans and Ignicoccus hospitalis: new insights into a unique, intimate association of two archaea. *J. Bacteriol.* **190**, 1743–1750 (2008).

184. E. Waters, M. J. Hohn, I. Ahel, D. E. Graham, M. D. Adams, M. Barnstead, K. Y. Beeson, L. Bibbs, R. Bolanos, M. Keller, K. Kretz, X. Lin, E. Mathur, J. Ni, M. Podar, T. Richardson, G. G. Sutton, M. Simon, D. Soll, K. O. Stetter, J. M. Short, M. Noordewier, The genome of Nanoarchaeum equitans: insights into early archaeal evolution and derived parasitism. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 12984–12988 (2003).

185. J. H. Munson-McGee, E. K. Field, M. Bateson, C. Rooney, R. Stepanauskas, M. J. Young, Nanoarchaeota, their Sulfolobales host, and Nanoarchaeota virus distribution across Yellowstone National Park Hot Springs. *Appl. Environ. Microbiol.* **81**, 7860–7868 (2015).

186. E. St. John, Y. Liu, M. Podar, M. B. Stott, J. Meneghin, Z. Chen, K. Lagutin, K. Mitchell, A.-L. Reysenbach, A new symbiotic nanoarchaeote (Candidatus Nanoclepta minutus) and its host (Zestosphaera tikiterensis gen. nov., sp. nov.) from a New Zealand hot spring. *Syst. Appl. Microbiol.* **42**, 94–106 (2019).

187. J. N. Hamm, S. Erdmann, E. A. Eloe-Fadrosh, A. Angeloni, L. Zhong, C. Brownlee, T. J. Williams, K. Barton, S. Carswell, M. A. Smith, S. Brazendale, A. M. Hancock, M. A. Allen, M. J. Raftery, R. Cavicchioli, Unexpected host dependency of Antarctic Nanohaloarchaeota. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 14661–14670 (2019).

188. S. Krause, A. Bremges, P. C. Münch, A. C. McHardy, J. Gescher, Characterisation of a stable laboratory co-culture of acidophilic nanoorganisms. *Sci. Rep.* **7**, 3289 (2017).

189. O. V. Golyshina, S. V. Toshchakov, K. S. Makarova, S. N. Gavrilov, A. A. Korzhenkov, V. La Cono, E. Arcadi, T. Y. Nechitaylo, M. Ferrer, I. V. Kublanov, Y. I. Wolf, M. M. Yakimov, P. N. Golyshin, 'ARMAN' archaea depend on association with euryarchaeal host in culture and in situ. *Nat. Commun.* **8** (2017).

190. N. H. Youssef, C. Rinke, R. Stepanauskas, I. Farag, T. Woyke, M. S. Elshahed, Insights into the metabolism, lifestyle and putative evolutionary history of the novel archaeal phylum "Diapherotrites." *ISME J.* **9**, 447–460 (2015).

191. J. T. Bird, B. J. Baker, A. J. Probst, M. Podar, K. G. Lloyd, Culture independent genomic comparisons reveal environmental adaptations for Altiarchaeales. *Front. Microbiol.* **7**, 1221 (2016).

192. R. Henneberger, C. Moissl, T. Amann, C. Rudolph, R. Huber, New insights into the lifestyle of the cold-loving SM1 euryarchaeon: natural growth as a monospecies biofilm in the subsurface. *Appl. Environ. Microbiol.* **72**, 192–199 (2006).

193. A. J. Probst, H.-Y. N. Holman, T. Z. DeSantis, G. L. Andersen, G. Birarda, H. A. Bechtel, Y. M. Piceno, M. Sonnleitner, K. Venkateswaran, C. Moissl-Eichinger, Tackling the minority: sulfate-reducing bacteria in an archaea-dominated subsurface biofilm. *ISME J.* **7**, 635–651 (2013).

194. C. Moissl, C. Rudolph, R. Huber, Natural communities of novel archaea and bacteria with a string-of-pearls-like morphology: molecular analysis of the bacterial partners. *Appl. Environ. Microbiol.* **68**, 933–937 (2002).

195. C. Rudolph, C. Moissl, R. Henneberger, R. Huber, Ecology and microbial structures of archaeal/bacterial strings-of-pearls communities and archaeal relatives thriving in cold sulfidic springs. *FEMS Microbiol. Ecol.* **50**, 1–11 (2004).

196. C. Moissl, C. Rudolph, R. Rachel, M. Koch, R. Huber, In situ growth of the novel SM1 euryarchaeon from a string-of-pearls-like microbial community in its cold biotope, its physical separation and insights into its structure and physiology. *Arch. Microbiol.* **180**, 211–217 (2003).

197. K. Schwank, T. L. V. Bornemann, N. Dombrowski, A. Spang, J. F. Banfield, A. J. Probst, An archaeal symbiont-host association from the deep terrestrial subsurface. *ISME J.* **13**, 2135–2139 (2019).

198. C. Moissl, R. Rachel, A. Briegel, H. Engelhardt, R. Huber, The unique structure of archaeal 'hami', highly complex cell appendages with nano-grappling hooks. *Mol. Microbiol.* **56**, 361–370 (2005).

199. T. L. Miller, M. J. Wolin, E. Conway de Macario, A. J. Macario, Isolation of Methanobrevibacter smithii from human feces. *Appl. Environ. Microbiol.* **43**, 227–232 (1982).

200. M. D. Levitt, J. K. Furne, M. Kuskowski, J. Ruddy, Stability of human methanogenic flora over 35 years and a review of insights obtained from breath methane measurements. *Clin. Gastroenterol. Hepatol.* **4**, 123–129 (2006).

201. N. Gaci, G. Borrel, W. Tottey, P. W. O'Toole, J.-F. Brugère, Archaea and the human gut: new beginning of an old story. *World J. Gastroenterol.* **20**, 16062–16078 (2014).

202. K. Koskinen, M. R. Pausan, A. K. Perras, M. Beck, C. Bang, M. Mora, A. Schilhabel, R. Schmitz, C. Moissl-Eichinger, First insights into the diverse human archaeome: Specific detection of Archaea in the gastrointestinal tract, lung, and nose and on skin. *MBio* **8** (2017).

203. M. R. Pausan, C. Csorba, G. Singer, H. Till, V. Schöpf, E. Santigli, B. Klug, C. Högenauer, M. Blohs, C. Moissl-Eichinger, Exploring the archaeome: Detection of Archaeal signatures in the human body. *Front. Microbiol.* **10**, 2796 (2019).

204. E. Conway de Macario, A. J. L. Macario, Methanogenic archaea in health and disease: a novel paradigm of microbial pathogenesis. *Int. J. Med. Microbiol.* **299**, 99–108 (2009).

205. Y. Sereme, S. Mezouar, G. Grine, J. L. Mege, M. Drancourt, P. Corbeau, J. Vitte, Methanogenic Archaea: Emerging partners in the field of allergic diseases. *Clin. Rev. Allergy Immunol.* **57**, 456–466 (2019).

206. T. Nguyen-Hieu, S. Khelaifia, G. Aboudharam, M. Drancourt, Methanogenic archaea in subgingival sites: a review. *APMIS* **121**, 467–477 (2013).

207. P. J. Pérez-Chaparro, C. Gonçalves, L. C. Figueiredo, M. Faveri, E. Lobão, N. Tamashiro, P. Duarte, M. Feres, Newly identified pathogens associated with periodontitis: a systematic review. *J. Dent. Res.* **93**, 846–858 (2014).

208. H. P. Horz, N. Robertz, M. E. Vianna, K. Henne, G. Conrads, Relationship between methanogenic archaea and subgingival microbial complexes in human periodontitis. *Anaerobe* **35**, 10–12 (2015).

209. H. P. Horz, G. Conrads, Methanogenic Archaea and oral infections - ways to unravel the black box. *J. Oral Microbiol.* **3**, 5940 (2011).

210. F. S. Ramiro, E. de Lira, G. Soares, B. Retamal-Valdes, M. Feres, L. C. Figueiredo, M. Faveri, Effects of different periodontal treatments in changing the prevalence and levels of Archaea present in the subgingival biofilm of subjects with periodontitis: A secondary analysis from a randomized controlled clinical trial. *Int. J. Dent. Hyg.* **16**, 569–575 (2018).

211. T. L. Miller, M. J. Wolin, Methanosphaera stadtmaniae gen. nov., sp. nov.: a species that forms methane by reducing methanol with hydrogen. *Arch. Microbiol.* **141**, 116–122 (1985).

212. H. P. Horz, G. Conrads, The discussion goes on: What is the role of Euryarchaeota in humans? *Archaea* **2010**, 967271 (2010).

213. K. F. Jarrell, A. D. Walters, C. Bochiwal, J. M. Borgia, T. Dickinson, J. P. J. Chong, Major players on the microbial stage: why archaea are important. *Microbiology* **157**, 919–936 (2011).

214. S. Khelaifia, D. Raoult, Haloferax massiliensis sp. nov., the first human-associated halophilic archaea. *New Microbes New Infect.* **12**, 96–98 (2016).

215. S. Khelaifia, A. Caputo, C. Andrieu, F. Cadoret, N. Armstrong, C. Michelle, J.-C. Lagier, F. Djossou, P.-E. Fournier, D. Raoult, Genome sequence and description of Haloferax massiliense sp. nov., a new halophilic archaeon isolated from the human gut. *Extremophiles* **22**, 485–498 (2018).

216. Y.-D. Nam, H.-W. Chang, K.-H. Kim, S. W. Roh, M.-S. Kim, M.-J. Jung, S.-W. Lee, J.-Y. Kim, J.-H. Yoon, J.-W. Bae, Bacterial, archaeal, and eukaryal diversity in the intestines of Korean people. *J. Microbiol.* **46**, 491–501 (2008).

217. P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, J. I. Gordon, The human microbiome project. *Nature* **449**, 804–810 (2007).

218. S. Khelaifia, M. Drancourt, Susceptibility of archaea to antimicrobial agents: applications to clinical microbiology. *Clin. Microbiol. Infect.* **18**, 841–848 (2012).

219. K. F. Jarrell, A. D. Walters, C. Bochiwal, J. M. Borgia, T. Dickinson, J. P. J. Chong, Major players on the microbial stage: why archaea are important (b). *Microbiology* **157**, 919–936 (2011).

220. B. Q. Minh, M. A. T. Nguyen, A. von Haeseler, Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).

221. S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).

CHAPTER 7

Synthesis
Synopsis and Outlook
Concluding Remarks

SYNTHESIS

The tree of life (TOL) is a useful framework from which to understand shared commonalities across the biosphere, the major transitions in cellular evolution, and the timing of key events in life's evolutionary history. Accurately resolving these periods in the planet's distant past are notoriously challenging, namely due to an incomplete fossil record with poor microbial representation and limitations in phylogenetic methodologies to properly resolve the deepest branches in the TOL. The research presented in this thesis has advanced our understanding of the TOL, including the root position, its major architecture, and the timing of major events in cellular evolution from the last universal common ancestor (LUCA) to extant organisms. A properly resolved and dated TOL was foundational for a renewed assessment of the evolutionary history of the ATP synthase. This work contributes novel phylogenetic dating techniques and highlights the importance of using robust and sophisticated bioinformatics methods which can be applied to larger, more diverse genomic datasets to properly resolve the early course of evolution.

RESOLVING THE DEEPEST SPLIT IN THE TREE OF LIFE IS SENSITIVE TO PHYLOGENETIC APPROACH

A resolved TOL underpins the debate over the root of the tree and the position of LUCA, the diversity of the two primary domains, and the events in the early course of evolution. Inferring such a tree has traditionally relied on a core set of universal marker genes, usually informational-processing genes, that resolve a long interdomain branch between Archaea and Bacteria (hereafter, AB branch). However, these universal marker genes represent a very small fraction of an organism's total genome, therefore raising speculation over whether this is sufficient information to adequately model phylogenetic relationships in the TOL (*1*). Recently, an expanded set of 381 marker genes was used to infer a phylogeny of over ten thousand Archaea and Bacteria, and resolved a shorter AB branch indicating a shorter divergence time between the two primary domains (*2*). The shape and branch lengths of the TOL inform evolutionary relationships, therefore such discrepancies in the AB branch length can hinder our understanding of the earliest periods of cellular evolution.

Our work reveals that the disparity in the AB branch length between this recently published study and traditional analyses was due to the inclusion of marker genes with complicated evolutionary histories that failed to recover the reciprocal monophyly of Archaea and Bacteria, in addition to poor model fit for the dataset. A fundamental condition for marker genes used in concatenated phylogenetic analysis is that their individual evolutionary histories must be consistent and congruent with that of the overall TOL, more specifically that the underlying single genes follow vertical patterns of evolution (*3–5*). Results presented here indicate that substantial horizontal gene transfer (HGT) between domains and hidden paralogy were major contributors to domain polyphyly and the artificial shortening of the AB branch. Recovery of domain or group monophyly is a fundamental condition that reliable marker genes must meet to be suitable for phylogenetic inference, because mixed phylogenetic signals can complicate

the final inference. We found that metabolic genes, and other non-ribosomal genes, in the expanded marker set (*2*) often failed to recover the reciprocal monophyly of Archaea and Bacteria. In addition to HGT, problematic markers showed evidence of extensive mixing of paralogous sequences with orthologous sequences. Importantly, we note that this was similarly evidenced in a smaller non-ribosomal marker gene set, that also resolved a relatively shorter AB branch length (*6*), highlighting the need to carefully select markers that reflect the evolutionary processes in the TOL or remove sequences with non-vertical signals prior to concatenation.

Our combined manual inspection and statistical examination of previously used ribosomal and non-ribosomal genes (*6–8*) indicated similar findings. Marker genes that failed to recover the reciprocal monophyly of the Archaea and Bacteria contributed to shortening of the AB branch. Taking it a step further, we applied a statistical ranking method, known as the *split score* (*9*), to identify suitable marker genes based on their ability to recover established taxonomic relationships. The inferred phylogeny from the 27 highest ranking marker genes identified the root of the TOL between the Archaea and Bacteria, which are separated by a long interdomain branch. Manual inspection stages in the phylogenetic workflow (see Chapter 2 Fig. 6) revealed that occasional HGT and paralogous mixing impacted the interdomain branch and should be removed for the confident inference of the TOL. Poor model fit was an added component of the phylogenetic approach that impacted the estimation of the AB branch length. Accounting for site-heterogeneity in the substitution model resulted in a longer inferred AB branch for both the top 27 best performing markers from our dataset and the original expanded marker gene set (*2*).

This work advances our understanding of concatenated phylogenetic inference of the TOL, demonstrating the need to carefully select marker genes, detect HGTs and paralogy, and determine the appropriate model fit for the data in order to reliably resolve deep evolutionary relationships. Results confirm that standard methods are reliable but require more advanced technical and methodological approaches that are more suitable for the wealth of biological information that is currently accessible and should be included in phylogenetic analyses. Traditional phylogenetic inference methods depending on a small set of universal marker genes are computationally tractable and a valuable window through which to explore evolutionary history. However, the inclusion of more marker genes and genomes is a valuable frontier to explore, and can enable a more comprehensive phylogenetic survey of evolutionary processes affecting the total genome. Moving forward, methods that account for and properly model both vertical and horizontal gene flow, such as the quantification of duplications, transfers, and losses across a genome using amalgamated likelihood estimation (ALE) (*10*, *11*), can serve as the foundation for more robust inference of the TOL.

USING ENDOSYMBIOSES TO DATE THE TOL

Having established the root position and shape of the TOL, this work continued on to tackle a more poorly modeled dimension of life's evolutionary history – its timing. Dating deep evolution in the prokaryotic lineages is extremely challenging due to an incomplete fossil record that is non-representative of the two primary domains and standard molecular clock methods

that struggle to effectively model variation in the rate of evolution along deep branches. Bypassing such critical limitations requires additional information, which can be accessed via the prokaryotic branches leading to eukaryotes in the TOL. The endosymbiosis events underlying eukaryogenesis involved an ancestral archaeal host and one or more bacterial partners. Therefore, eukaryotes can have between 2-3 distinct genetic sources of certain proteins and machinery: a nuclear source from the archaeal host, a mitochondrial source, and a plastid source. Integrating this eukaryotic information into phylogenetic analyses including Archaea and Bacteria results in multiple nodes across the respective tree corresponding to the same evolutionary event. Equivalent nodes can be *cross-braced* (*12*, *13*) to the same fossil age to apply an additional constraint on the relative rates of evolution along intervening branches. This cross-bracing approach was combined with a relative time constraint (*14*) that positions nodes by relative age or emergence in the tree, for example, the mitochondrial endosymbiosis must be older than the plastid endosymbiosis. This approach enabled us to propagate the eukaryotic fossil evidence into the deeper regions of both primary domains.

This newly devised dating technique was applied to a TOL including Archaea, Bacteria, and three eukaryotic sources (nuclear, mitochondrial, and plastid) to time cellular evolution using a series of known fossil calibrations (Chapter 3 Fig. 5A, C). Results showed that LUCA and LBCA lived in overlapping time periods relatively proximal to the moon-forming impact, from 4.52-4.32 Ga and 4.49-4.05 Ga, respectively. Both LUCA and LBCA appeared to predate the last archaeal common ancestor (LACA, 3.95-3.37 Ga). This is particularly interesting because this suggests that the ancestor of extant archaea radiated after the Bacteria, as it is comparable in age to some of the deepest bacterial lineages. However, this could also be the result of a sampling bottleneck along the archaeal stem. The endosymbiosis leading to mitochondria was estimated to have occurred around 2.58-2.12 Ga, which overlaps with the timing of the last ancestor between the nuclear eukaryotic lineage and its closest asgardarchaeal relative (2.67-2.19 Ga). While the nuclear and mitochondrial LECA nodes can be cross-braced to the same age, the preceding nuclear lineage stem was moderately longer than the mitochondrial lineage stem, suggesting divergence from the most recent asgardarchaeal ancestor occurred before the divergence of mitochondria from the alphaproteobacterial ancestor. This finding is compatible with an intermediate or late acquisition of the mitochondria, consistent with recent work on the timing of mitochondrial acquisition (*15–17*). However, it is difficult to determine whether other factors are impacting the variation in stem length, such as differences in relative evolutionary rates in host versus mitochondrial-derived genes. Furthermore, while these divergences outline important questions surrounding the timing of mitochondrial acquisition relative to other eukaryotic features, they do not allow us to directly test these hypotheses as discussed in our manuscript.

While the Cyanobacteria radiated around 2.74-2.03 Ga, the plastid origin was dated to between 2.14-1.73 Ga. The age of the last eukaryotic common ancestor (LECA, 1.93-1.84 Ga) is compatible with previously proposed scenarios (*18–22*), falling into an interval spanning ~2.4-1.0 Ga. This dated TOL provides a timeline of the evolution of other key taxonomic groups of interest and anchors enzyme evolution through time. We extended this principle to provide an absolute timeline for ATP synthase evolution.

A TIMELINE FOR ATP SYNTHASE ORIGINS AND A REVISED EVOLUTIONARY SCENARIO

The ATP synthase is a central energy generating enzymatic complex found across all domains of life. Functionally, the ATP synthase mediates the production of ATP, which is often regarded as the energy currency of life. Due to its highly conserved nature across the biosphere, it has long been speculated to have evolved very early in the timeline of cellular evolution, with some proposing its origination during the pre- to proto-cell transition period (*23–25*). The ATP synthase has been instrumental in phylogenetic analyses due to the ancient gene duplication of the key catalytic subunit in the rotary headpiece of the complex. This very ancient gene duplication followed by a loss of function in one of the duplicated products resulted in the characteristic heterohexameric headpiece of this enzyme that is composed of triplicates of a catalytic and so-called non-catalytic subunit, which together form the site of ATP synthesis. These paralogs can each serve as a distant outgroup for the other providing a means to root the TOL (*26–28*).

The diversification of the ATP synthase was presumed to align with the diversification of the three domains of life, with the F-, A-, and V-types being native to Bacteria, Archaea, and eukaryotes, respectively. Early phylogenetic analyses into eukaryotic ATP synthases demonstrated the horizontal inheritance pattern of this enzymatic complex, paralleling the endosymbiosis events that gave rise to the protoeukaryote (mitochondrial inheritance) and the photosynthetic eukaryotes (plastid inheritance) (*29–31*). Specifically, the so-called V-type ATP synthase in eukaryotes appeared to be inherited from the ancestral archaeal host, whereas the F-type ATP synthases in the mitochondria and plastids of eukaryotes descend from their bacterial ancestors, the Alphaproteobacteria and Cyanobacteria, respectively. The diversification of the F- and A/V-type (a larger family consisting of both the A- and V-type complexes) ATP synthases have traditionally been consistent with a root of the TOL between Archaea and Bacteria. A/V-type complexes initially seemed patchily distributed across the Bacteria, which was proposed to be the result of horizontal transfer of genes from Archaea to Bacteria inhabiting shared environments (*25*, *30*, *32*).

Similar to the ribosomal proteins, endosymbioses underlie the evolutionary processes in ATP synthase gene trees of the catalytic and non-catalytic subunits, with eukaryotes containing up to three different ATP synthases (*27*, *29–31*). The ancient pre-LUCA gene duplication results in speciation events appearing at least twice in the ATP synthase gene tree of the catalytic and non-catalytic subunits. In turn, we could apply the same cross-bracing technique used to time the TOL to date key events in ATP synthase evolution such as the ancient duplication of the catalytic subunits (i.e., the split between the catalytic and non-catalytic subunits) and diversification into the F- and A/V-type ATP synthases. Our results show that the split between the catalytic and non-catalytic subunits occurred from 4.52-4.46 Ga, possibly predating or overlapping with LUCA, and consistent with the possible primordial evolution of the complex. In addition, our analyses revealed that the diversification of the catalytic and non-catalytic subunits of the F- and A/V-type ATP synthases also occurred very early, dating to 4.52-4.38 Ga and 4.52-4.42 Ga, respectively.

Timing the diversification of the F- and A/V-type ATP synthases to a period overlapping with LUCA and LBCA but preceding LACA invites an updated scenario to explain ATP synthase evolution. It is possible that diversification of the F- and A/V-type ATP synthases preceded the diversification of the Archaea and Bacteria, with both types already present in LUCA. In this scenario, the F-type ATP synthases are predicted to be lost along the stem to LACA, while both complexes were inherited by LBCA (Chapter 3 Fig. 4C). This finding is consistent with the inference of the F- and A/V-type ATP synthase to the proteome of LBCA (Chapter 4), the prevalence of bacterial A/V-type complexes, and paucity of archaeal F-type complexes (Chapter 3 Fig. 1). However, we cannot rule out an alternative scenario where a primitive F-type-like ATP synthase diverged into the two major types along their respective stems to LACA and LBCA, and an evolved A/V-type was transferred into the stem to LBCA (Chapter 3 Fig. 4B).

While this work provides immense insight into ATP synthase evolution, the timing of cellular evolution, and the phylogenetic and molecular clock technologies to conduct such analyses, it opened many unanswered questions. Deeper sampling of Archaea would provide better insight into whether the younger age for LACA is a phylogenetic artifact from a sequencing bottleneck along the archaeal stem. It would be particularly interesting to examine the ATP synthases of newly defined lineages in the Asgardarchaeota and DPANN, considering the important implications they have for understanding cellular evolution. Assessing the F- and A/V-type ATP synthases in recently defined archaeal and bacterial lineages could provide greater insight into the distribution of these complexes along the TOL and provide additional information from which to address the mechanisms underlying the prevalence of bacterial A/V-types compared to their archaeal F-type counterparts. This is especially intriguing when considering that HGT from Bacteria to Archaea appears to be more common (*33*). While cross-bracing with a relative constraint on trees with equivalent nodes was effective to time cellular evolution, some of the oldest estimated ages in the TOL and ATP synthase evolution remain to be validated. Specifically, many of the cross-braced age ranges approach the maximum fossil age that we applied for the root node: 4.52 Ga, corresponding to the moon-forming impact in the late Hadean period (*18*). Immense debate exists over the origins of life around this time period, as environments on the early Earth are predicted to have experienced many physical and chemical extremes, and have been proposed to be incompatible with life. A shielded start to life at submarine hydrothermal vent systems may have been a solution to the exposure to environmental extremes near the surface. Nonetheless, the earliest periods of planetary evolution following the moon-forming impact could have been marked by rapid and frequent evolutionary innovation (*34*), and therefore LUCA appearing during this period would not be impossible. Additional prokaryotic and eukaryotic fossils could help address this issue by providing more time calibrations to the molecular dating analysis. A major issue with identifying fossils in the earliest periods of planetary and biological evolution is that during the late Hadean, sedimentary rock did not deposit as frequently in the superhot environment. Similarly, since prokaryotes are microscopic, it can be very difficult for them to leave fossil deposits behind unless they are involved in a biogeochemistry that can be fossilized in sediments without major disruption. A promising alternative to physical prokaryotic fossils

could be the identification of additional indirect markers of metabolic activity of life. For example, the major geologic changes that occurred during the Great Oxidation Event (GOE) around 2.3 Ga, suggest that oxygenic photosynthesis, a metabolism specific to Cyanobacteria, emerged prior to this event. Using such indirect markers, Davín and coworkers developed a timescale for bacterial evolution using aerobic and anaerobic inventions linked to the geochemical record, specifically calibrated to the GOE (*35*). Consistent with findings presented here, they dated early bacterial radiation to the Archaean. Identification of other metabolic signatures from prokaryotes could provide important calibrations to further improve molecular dating in the future.

BACTERIAL EVOLUTION IS TREELIKE AND LBCA WAS AN ALREADY COMPLEX FREE-LIVING CELL

The interplay between vertical and horizontal evolution is challenging to quantify and model using standard phylogenetic methods. Results presented in Chapter 2 indicate that failure to properly account for gene exchange, such as interdomain transfer, can impact the shape and topology of the TOL. It was previously noted that using a small subset of marker genes does not effectively capture all evolutionary processes at play in the total genome of an organism (*1*). Another common limitation to phylogenetic inference is reliance on outgroup-rooting, which utilizes a distantly related group to better resolve the within-group phylogenetic relationships. In some cases, gene sequences with compositional biases, such as those from organisms with reduced genomes or that undergo accelerated evolution, can erroneously be attracted to one another and toward the distant outgroup, a phenomenon known as long branch attraction (LBA) (*36*, *37*). Accounting for such phylogenetic artifacts is crucial to properly place potentially fast evolving or long-branching lineages in their respective phylogenies. An alternative solution to root phylogenetic trees without a distant outgroup is to use phylogenetic tools that model the evolution of all protein families across a taxon set and reconcile those with a species tree. Specifically, gene-tree aware methods model the evolutionary history of all individual gene trees along a rooted species tree of those taxa, quantifying gene duplications, loss, originations and transfers (DTOL) between genomes. The application of gene tree-species tree reconciliation methods (*10*, *38*) extends beyond a means to statistically test root positions and can also be applied to model gene flow. Specifically, since ALE is a probabilistic approach, it estimates the probability of each gene experiencing DTOLs at each node position on the tree. These probabilities can then be used to determine the enzymatic repertoire of key ancestors in the tree.

Using this gene tree-species tree reconciliations approach, we resolved the root of the bacterial phylogeny to be between two major clades, the Terrabacteria and Gracilicutes with Fusobacteria supported on either side of the root (Chapter 4 Fig. 1A and 1B). Quantification of the DTOLs revealed that bacterial genomes have a strong vertical component of evolution, with the majority of genes descending vertically (66%) compared to horizontally (34%), making a tree a reliable framework based on which to visualize bacterial evolution. We resolved the placement of the candidate phyla radiation (CPR), a large early-branching bacterial radiation of

putative symbionts with compositionally biased genomes (*39*, *40*). Previous analyses positioned the CPR at the base of the bacterial clade (*2*, *39–42*), however our analysis places them sister to the Chloroflexota as a derived lineage within the Terrabacteria, consistent with findings from the phylogenetic inference in Chapter 2, and other more recent evidence (*43*, *44*). CPR as a derived lineage implies that they evolved by genome reduction from a free-living ancestor.

Reconstruction of LBCA's proteome revealed a complex rod-shaped, free-living ancestor that was likely motile with the ability to sense environmental stimuli. Results indicated high support for didermy and the presence of genes that anchor flagella and pili in double membranes. Didermy in the ancestor of bacteria implies monoderms evolved on several separate occasions (*45*, *46*). LBCA appears to have relied on glycolysis, the tricarboxylic acid cycle (TCA), and the pentose-phosphate pathway (PPP) for carbohydrate metabolism. However, recovery of genes associated with carbon-fixation was patchier; we detected support for some genes in the TCA and reductive glycine pathways, but it was not possible to ascertain directionality of the reactions. The Wood-Ljungdahl pathway (WLP) is believed to be one of the most ancient microbial metabolisms, possibly emerging proximal to the origins of cellular life itself (*23*, *47*). We recovered high support for several enzymes belonging to the methyl branch (acetogenic bacteria) of the WLP, but only moderate support for the central key enzyme of this pathway, the carbon monoxide dehydrogenase acetyl coenzyme-A synthase (CODH-ACS). However, when considered in conjunction with the presence of enzymes belonging to the Rhodobacter nitrogen-fixation (RNF) complex, it is likely that LBCA was capable of acetogenic growth. We detected a near-complete CRISPR-Cas system, suggesting the bacterial ancestor evolved in an environment where it was exposed to viruses and other parasitic replicators. The presence of both the F- and A/V-type ATP synthases in LBCA's proteome is crucial to understanding alternative scenarios of ATP synthase evolution and in line with our findings on ATP synthase evolution (Chapter 3).

Advanced techniques such as those applied here can address some of the largest open questions in deep evolution while overcoming the limitations of standard phylogenetic methods. Outgroup-free rooting of the archaeal domain will be useful in examining the position of the so-called DPANN, another early-branching and diverse lineage. The DPANN are often resolved as an early diverging branch in the archaeal tree (*9*, *40*, *48–51*), which by some, has been attributed to represent a potential phylogenetic artifact (*52*, *53*). Analyzes with a larger archaeal dataset in combination with more sophisticated phylogenetic approaches can help confirm or refute these phylogenetic proposals. Inferring the proteome of the major ancestors across the TOL, including LUCA, could shed light into metabolic evolution from life's earliest origins to the major domains.

SYNOPSIS AND OUTLOOK

Overall, the results presented in this thesis demonstrate that advanced phylogenetic methods are required to gain greater insight into deep evolutionary events. This work addressed key questions regarding the shape of the TOL including its root position, the diversification of the primary domains, the genetic contributions and timing of eukaryogenesis, the timing of cellular evolution, and the evolution of key enzymes and metabolic processes. Specifically, my work has provided following novel key insights into early cellular evolution and phylogenetic approaches:

1) Marker gene selection and model fit can greatly impact phylogenetic inference: Failure to account for frequent HGT and paralogous families result in single gene phylogenies that are incongruent with the overall TOL and can artificially draw the two major domains closer together. Careful manual inspection is required to detect and filter out any major cases of HGT and paralogous families.

2) The phylogenetic implications of endosymbioses and ancient gene duplications: Certain enzymatic machinery was inherited into eukaryotes from their distinct prokaryotic ancestors. Phylogenetically, this results in node equivalence across a species or gene tree, which can be used to constrain fossil calibrations for molecular dating.

3) The early planetary timeline was a period of rapid innovation: LUCA and LBCA emerged soon after the formation of the Earth, suggesting the earliest stages in the biological timeline were periods of rapid evolutionary innovation.

4) ATP synthases occurred very early in the evolutionary timeline: The split between the catalytic and non-catalytic subunit and the diversification of the F- and A/V-type ATP synthases occurred in a period overlapping with LUCA and LBCA.

5) Bacterial evolution is treelike and is affected by both vertical and horizontal gene flow: Sophisticated phylogenomic methods that model gene tree evolution against a species tree resolved a bacterial tree split between the Terrabacteria and Gracilicutes. Additionally, quantification of the duplications, transfers, originations, and losses revealed that most genes are subject to HGT at some point in their history, but the majority of genes are transmitted vertically.

6) LBCA was a complex ancestor that interacted with its environment: The proteomic reconstruction highlighted the complexity of LBCA's metabolic repertoire, with evidence of a complex diderm ancestor capable of sensing its environment.

7) Viruses are important to the earliest periods of cellular evolution: A CRISPR-Cas system was resolved in the ancestral genome reconstructions of LBCA and LUCA (see Appendix A), suggesting viruses and other parasitic replicators were coevolving alongside cellular life from its earliest periods.

Methods that were tested and developed in these projects hold promise in enhancing phylogenetic approaches that can model deep evolutionary events. However, open questions remain regarding the diversity of the archaeal domain and key metabolic inventions from the earliest periods of cellular evolution.

Gene tree-species tree reconciliation techniques (Chapter 4) should be applied to an even greater sampling of Archaea to properly address the root position of the domain, the placement of the DPANN lineage, and the metabolic potential of LACA. In ongoing work, I have started to explore the evolutionary history of the Wood-Ljungdahl pathway in Archaea to address key questions about archaeal evolution and metabolic history. The WLP has been proposed to be one of the earliest metabolisms in the evolution of cellular life (*54*). Of the known carbon-fixation pathways, the WLP is the only one that produces ATP (*55*). This is achieved through the redox reaction involving H_2 and CO_2 and mediated by iron-sulfur clusters, all of which are abundant at alkaline hydrothermal vents. Previous analyses have inferred LUCA and LACA to contain the WLP pathway for carbon-fixation (*24*, *51*). Until recently, Altiarchaeota were the only DPANN lineage containing the WLP (*56*, *57*), but their placement is unstable and therefore complicates tracing the trajectory of the WLP across the DPANN. Research has uncovered two distinct taxa belonging to the undescribed SpSt_1190 lineage (*58*) in a marine oxygen deficient zone. Our preliminary phylogenetic analyses found that the SpSt_1190 lineage branches within the highly-reduced Cluster 2 DPANN (*9*). Interestingly, the SpSt_1190 lineage contains two representatives, SpSt_1190_A and SpSt_1190_B (Fig. 1), which have a reduced genome of approximately 1Mb and an average genome of 4Mb, respectively, the latter of which encodes all the genes for the WLP and other auxiliary genes for acetogenic growth. Novel members of the SpSt lineage have been identified in samples from Black Sea metagenomes, which has allowed me to resolve a new tree of Archaea (Fig. 1), considering expanded archaeal diversity. Inspection of the SpSt_1190 genomes indicated a presence of all WLP genes in SpSt_1190_B-NIOZ117_mb_b257_2 and three subunits in SpSt_1190_B-NIOZ118_bs_b358_2 (Fig. 1). Initial findings demonstrate a marked difference in the distribution of WLP genes across the SpSt, with SpSt_1190_B containing most genes associated with the key metabolic enzyme, the CODH/ACS. In addition, both members of SpSt_1190_B contain a full Rhodobacter nitrogen fixation (Rnf) complex (*59*) (not shown), which enables the formation of a transmembrane ion gradient by coupling the oxidation of ferredoxin to the reduction of NAD+ (*60*). Redox intermediates produced via the Rnf complex can be fed into the WLP. Together, these findings suggest these organisms may be capable of acetogenic growth. This preliminary data indicates that there are distinct metabolic differences between the two clans of the SpSt_1190 lineage. To examine this further, I plan to develop a metabolic profile for each group (i.e., SpSt_1190_A and SpSt_1190_B) and use gene tree-species tree reconciliations to infer the origination of the WLP genes across the archaeal phylogeny. Furthermore, I will apply phylogenetic and comparative genomic analyses to assess whether members of the SpSt_1190 are host-dependent or free living.

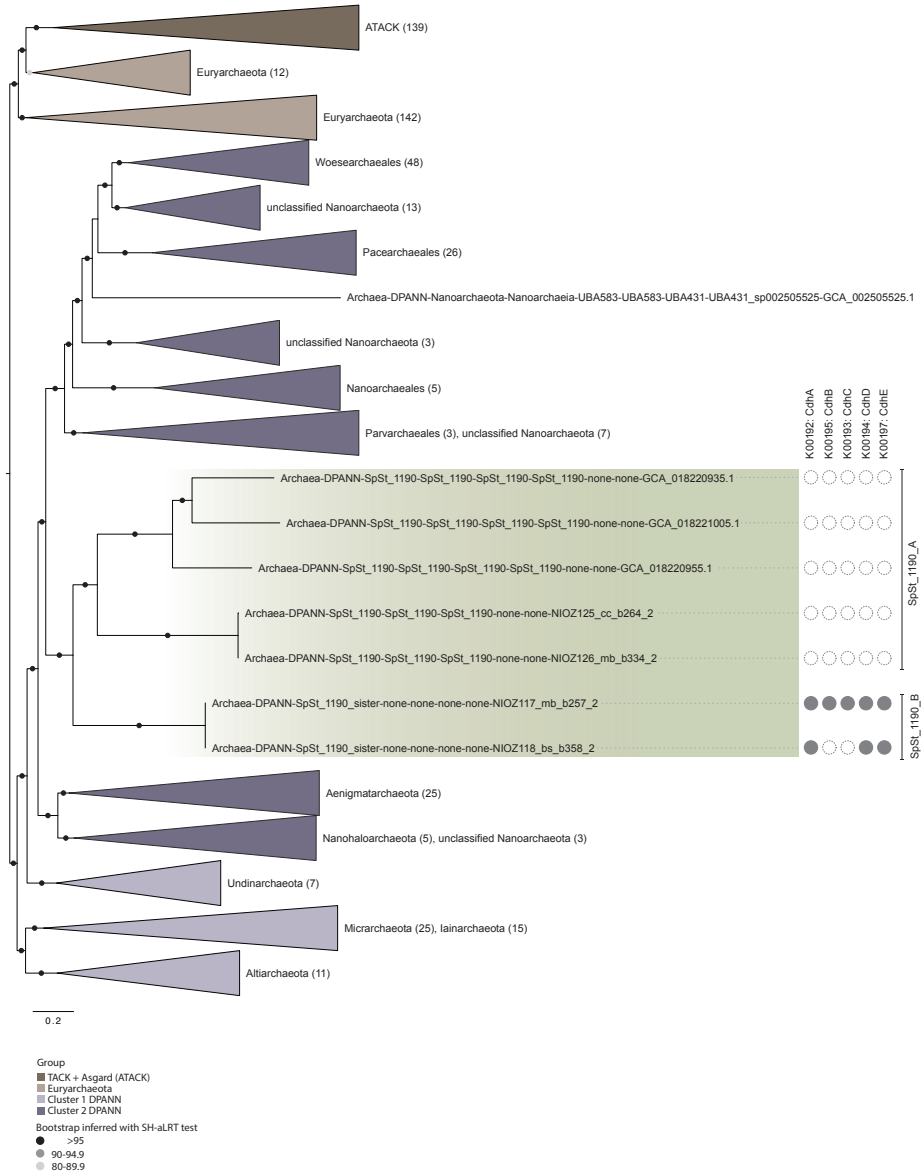


Fig. 1 | Distribution of the key subunits of the carbon monoxide dehydrogenase acetyl CoA synthase mapped to the SpSt_1190 lineage. COG families representing the key subunits of the CODH/ACS (CdhABCDE) were identified on the genomes of the SpSt_1190_A and SpSt_1190_B clans of the SpSt_1190 lineage. Dark shaded circles indicate presence, unshaded circles indicate absence. The concatenated alignment contains 497 taxa, was computed using MAFFT-LiNSi v7.453 (settings: --reorder) (*61*) and trimmed with BMGE v1.12 (settings: BLOSUM30 -h 0.55) (*62*). The final concatenated alignment contained 11,288 sites. The maximum-likelihood tree was inferred using IQ-TREE2 v2.1.2 with the LG+C60+R+F model (the best-fit model chosen by BIC) and SH-like approximate likelihood (see legend) and ultrafast bootstrap approximation (not shown), each with 1000 replicates (*63–65*). Scale bar represents the expected number of substitutions per site. Color code: ATACK (Asgard + TACK Archaea) = dark brown, Euryarchaeota = light brown, Cluster 1 DPANN = light purple, Cluster 2 DPANN = dark purple.

We also applied probabilistic reconciliations techniques to an evenly sampled set of archaeal and bacterial genomes to reconstruct the proteome of LUCA and time its evolution (Appendix A). Results indicate that LUCA was likely an anaerobe with genes for both the archaeal and bacterial branches of the WLP, suggesting potential for acetogenic growth. It is inconclusive whether LUCA was a heterotroph or autotroph, with results indicating either lifestyle is possible based on the presence of genes belonging to glycolysis, gluconeogenesis, the WLP, and the pentose-phosphate-pathway. Findings suggest that LUCA was already a complex cell, resembling extant Archaea and Bacteria, but details of its membrane composition is still debated. Dating analyses infer LUCA to have lived approximately 4.2 Ga, which is consistent with our inference of LUCA's age using our cross-bracing approach (Chapter 3).

Studies detailed here underscore the significance of investigating the evolutionary history of viruses (Chapter 5). Both LBCA (Chapter 4) and LUCA (Appendix A) were inferred to contain CRISPR-Cas systems, indicating that early cellular life was co-evolving with viruses or other parasitic replicators. Viruses are known to be integral to horizontal transfer of genetic material between organisms across the TOL. Additionally, recent phylogenetic analyses have revealed that LUCA had a diverse and complex virome (*66*), and that the genetic module may have originated in the primordial replicon pool, whereas the capsid components may have been acquired on different occasions from different cellular hosts (*67*). Nevertheless, viral evolution and indicators that cellular ancestors were equipped with viral defense mechanisms would suggest they were prevalent in the early Earth environment. The frontier of this work would be to increase sampling of viral genomes and conduct more detailed phylogenetic assessments of viruses in the context of the TOL.

CONCLUDING REMARKS

In this thesis, I applied sophisticated phylogenetic approaches to disentangle the most enigmatic and poorly understood aspects of early cellular evolution and diversification across the TOL. Cumulatively, this work has advanced our knowledge of the shape and structure of the TOL from the deepest split at the root to the prokaryotic branches involved in eukaryogenesis. This thesis highlights the importance of using robust and diverse phylogenetic approaches to address open questions regarding cellular and enzymatic evolution. A central theme of this thesis highlights how crucial and necessary advances in technology and technical approaches are in probing complex evolutionary relationships, especially when considering large expanses of biodiversity that have been discovered in recent years. Research reported in this thesis puts emphasis on how modeling enzyme evolution at different levels, from single enzymatic complexes to modeling all of the protein families in a dataset, can provide valuable insight into key evolutionary questions. Analysis of various marker gene approaches uncovered the need to carefully inspect and ensure the necessary criteria are met for markers used to infer concatenated phylogenies (Chapter 2). Beyond that, we observed that enzymes that are transmitted from prokaryotes to eukaryotes by endosymbiosis, are a useful window into

deep evolutionary events in prokaryotes (Chapter 3). Modeling the evolutionary histories of all proteins in a dataset is computationally demanding but provides a foundation for addressing a multitude of questions, including the treelike nature of evolution when considering the proportion of vertical and horizontal gene flow acting on a dataset. Furthermore, through modeling all protein families, one can address otherwise difficult questions regarding the origination/presence of genes across a phylogeny (Chapter 4).

While work reported here has contributed substantially to the field of phylogenetics and evolutionary biology, more questions have emerged over what additional techniques can be developed or applied to such analyses to refine our interpretation of deep evolution in the future. In general, the granularity of phylogenetic relationships is enhanced with the inclusion of more genomic information, however the tradeoff is the increased computational burden for such analyses. Exploring methods that can strike a balance between these two principles could greatly benefit such studies (Chapter 5).

REFERENCES

1. T. Dagan, W. Martin, The tree of one percent. *Genome Biol.* **7**, 118 (2006).

2. Q. Zhu, U. Mai, W. Pfeiffer, S. Janssen, F. Asnicar, J. G. Sanders, P. Belda-Ferre, G. A. Al-Ghalith, E. Kopylova, D. McDonald, T. Kosciolek, J. B. Yin, S. Huang, N. Salam, J.-Y. Jiao, Z. Wu, Z. Z. Xu, K. Cantrell, Y. Yang, E. Sayyari, M. Rabiee, J. T. Morton, S. Podell, D. Knights, W.-J. Li, C. Huttenhower, N. Segata, L. Smarr, S. Mirarab, R. Knight, Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477 (2019).

3. D. L. Theobald, A formal test of the theory of universal common ancestry. *Nature* **465**, 219–222 (2010).

4. P. Puigbò, Y. I. Wolf, E. V. Koonin, Search for a "Tree of Life" in the thicket of the phylogenetic forest. *J. Biol.* **8**, 59 (2009).

5. F. D. Ciccarelli, T. Doerks, C. von Mering, C. J. Creevey, B. Snel, P. Bork, Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287 (2006).

6. C. Petitjean, P. Deschamps, P. López-García, D. Moreira, Rooting the domain archaea by phylogenomic analysis supports the foundation of the new kingdom Proteoarchaeota. *Genome Biol. Evol.* **7**, 191–204 (2014).

7. T. A. Williams, C. J. Cox, P. G. Foster, G. J. Szöllősi, T. M. Embley, Phylogenomics provides robust support for a two-domains tree of life. *Nat Ecol Evol* **4**, 138–147 (2020).

8. G. A. Coleman, A. A. Davín, T. A. Mahendrarajah, L. L. Szánthó, A. Spang, P. Hugenholtz, G. J. Szöllősi, T. A. Williams, A rooted phylogeny resolves early bacterial evolution. *Science* **372** (2021).

9. N. Dombrowski, T. A. Williams, J. Sun, B. J. Woodcroft, J.-H. Lee, B. Q. Minh, C. Rinke, A. Spang, Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. *Nat. Commun.* **11**, 3939 (2020).

10. G. J. Szöllősi, W. Rosikiewicz, B. Boussau, E. Tannier, V. Daubin, Efficient exploration of the space of reconciled gene trees. *Syst. Biol.* **62**, 901–912 (2013).

11. B. Morel, P. Schade, S. Lutteropp, T. A. Williams, G. J. Szöllősi, A. Stamatakis, SpeciesRax: A Tool for Maximum Likelihood Species Tree Inference from Gene Family Trees under Duplication, Transfer, and Loss. *Mol. Biol. Evol.* **39** (2022).

12. P. M. Shih, N. J. Matzke, Primary endosymbiosis events date to the later Proterozoic with cross-calibrated phylogenetic dating of duplicated ATPase proteins. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 12355–12360 (2013).

13. P. P. Sharma, W. C. Wheeler, Cross-bracing uncalibrated nodes in molecular dating improves congruence of fossil and molecular age estimates. *Front. Zool.* **11**, 1–13 (2014).

14. G. J. Szöllősi, S. Höhna, T. A. Williams, D. Schrempf, V. Daubin, B. Boussau, Relative Time Constraints Improve Molecular Dating. *Syst. Biol.* **71**, 797–809 (2022).

15. A. A. Pittis, T. Gabaldón, Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature* **531**, 101–104 (2016).

16. T. J. G. Ettema, Evolution: Mitochondria in the second act, *Nature*. **531** (2016)pp. 39–40.

17. J. Vosseberg, J. J. E. van Hooff, M. Marcet-Houben, A. van Vlimmeren, L. M. van Wijk, T. Gabaldón, B. Snel, Timing the origin of eukaryotic cellular complexity with ancient duplications. *Nat Ecol Evol* **5**, 92–100 (2021).

18. H. C. Betts, M. N. Puttick, J. W. Clark, T. A. Williams, P. C. J. Donoghue, D. Pisani, Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nat Ecol Evol* **2**, 1556–1562 (2018).

19. L. W. Parfrey, D. J. G. Lahr, A. H. Knoll, L. A. Katz, Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 13624–13629 (2011).

20. L. Eme, S. C. Sharpe, M. W. Brown, A. J. Roger, On the age of eukaryotes: evaluating evidence from fossils and molecular clocks. *Cold Spring Harb. Perspect. Biol.* **6** (2014).

21. C. Berney, J. Pawlowski, A molecular time-scale for eukaryote evolution recalibrated with the continuous microfossil record. *Proc. Biol. Sci.* **273**, 1867–1872 (2006).

22. D. Chernikova, S. Motamedi, M. Csürös, E. V. Koonin, I. B. Rogozin, A late origin of the extant eukaryotic diversity: divergence time estimates using rare genomic changes. *Biol. Direct* **6**, 26 (2011).

23. N. Lane, J. F. Allen, W. Martin, How did LUCA make a living? Chemiosmosis in the origin of life. *Bioessays* **32**, 271–280 (2010).

24. M. C. Weiss, F. L. Sousa, N. Mrnjavac, S. Neukirchen, M. Roettger, S. Nelson-Sathi, W. F. Martin, The physiology and habitat of the last universal common ancestor. *Nat Microbiol* **1**, 16116 (2016).

25. A. Y. Mulkidjanian, M. Y. Galperin, K. S. Makarova, Y. I. Wolf, E. V. Koonin, Evolutionary primacy of sodium bioenergetics. *Biol. Direct* **3**, 13 (2008).

26. N. Iwabe, K. Kuma, M. Hasegawa, S. Osawa, T. Miyata, Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 9355–9359 (1989).

27. J. P. Gogarten, H. Kibak, P. Dittrich, L. Taiz, E. J. Bowman, B. J. Bowman, M. F. Manolson, R. J. Poole, T. Date, T. Oshima, J. Konishi, K. Denda, M. Yoshida, Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 6661–6665 (1989).

28. C. R. Woese, O. Kandler, M. L. Wheelis, Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 4576–4579 (1990).

29. J. P. Gogarten, L. Taiz, Evolution of proton pumping ATPases: Rooting the tree of life. *Photosynth. Res.* **33**, 137–146 (1992).

30. E. Hilario, J. P. Gogarten, Horizontal transfer of ATPase genes--the tree of life becomes a net of life. *Biosystems.* **31**, 111–119 (1993).

31. E. Hilario, J. P. Gogarten, The prokaryote-to-eukaryote transition reflected in the evolution of the V/F/A-ATPase catalytic and proteolipid subunits. *J. Mol. Evol.* **46**, 703–715 (1998).

32. P. Lapierre, R. Shial, J. P. Gogarten, Distribution of F- and A/V-type ATPases in Thermus scotoductus and other closely related species. *Syst. Appl. Microbiol.* **29**, 15–23 (2006).

33. S. Nelson-Sathi, F. L. Sousa, M. Roettger, N. Lozada-Chávez, T. Thiergart, A. Janssen, D. Bryant, G. Landan, P. Schönheit, B. Siebers, J. O. McInerney, W. F. Martin, Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* **517**, 77–80 (2015).

34. L. A. David, E. J. Alm, Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* **469**, 93–96 (2011).

35. A. A. Davín, B. J. Woodcroft, R. M. Soo, B. Morel, R. Murali, D. Schrempf, J. Clark, B. Boussau, E. R. R. Moody, L. L. Szánthó, E. Richy, D. Pisani, J. Hemp, W. Fischer, P. C. J. Donoghue, A. Spang, P. Hugenholtz, T. A. Williams, G. J. Szöllősi, An evolutionary timescale for Bacteria calibrated using the Great Oxidation Event, *bioRxiv* (2023)p. 2023.08.08.552427.

36. J. Bergsten, A review of long-branch attraction. *Cladistics* **21**, 163–193 (2005).

37. L. Shavit, D. Penny, M. D. Hendy, B. R. Holland, The problem of rooting rapid radiations. *Mol. Biol. Evol.* **24**, 2400–2411 (2007).

38. G. J. Szöllosi, B. Boussau, S. S. Abby, E. Tannier, V. Daubin, Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17513–17518 (2012).

39. C. T. Brown, L. A. Hug, B. C. Thomas, I. Sharon, C. J. Castelle, A. Singh, M. J. Wilkins, K. C. Wrighton, K. H. Williams, J. F. Banfield, Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).

40. L. A. Hug, B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D. A. Relman, K. M. Finstad, R. Amundson, B. C. Thomas, J. F. Banfield, A new view of the tree of life. *Nat Microbiol* **1**, 16048 (2016).

41. C. J. Castelle, J. F. Banfield, Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. *Cell* **172**, 1181–1197 (2018).

42. R. Méheust, D. Burstein, C. J. Castelle, J. F. Banfield, The distinction of CPR bacteria from other bacteria based on protein family content. *Nat. Commun.* **10**, 4173 (2019).

43. N. Taib, D. Megrian, J. Witwinowski, P. Adam, D. Poppleton, G. Borrel, C. Beloin, S. Gribaldo, Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. *Nat Ecol Evol* **4**, 1661–1672 (2020).

44. C. A. Martinez-Gutierrez, F. O. Aylward, Phylogenetic Signal, Congruence, and Uncertainty across Bacteria and Archaea. *Mol. Biol. Evol.* **38**, 5514–5527 (2021).

45. D. Megrian, N. Taib, J. Witwinowski, C. Beloin, S. Gribaldo, One or two membranes? Diderm Firmicutes challenge the Gram-positive/Gram-negative divide. *Mol. Microbiol.* **113**, 659–671 (2020).

46. J. Witwinowski, A. Sartori-Rupp, N. Taib, N. Pende, T. N. Tham, D. Poppleton, J.-M. Ghigo, C. Beloin, S. Gribaldo, An ancient divide in outer membrane tethering systems in bacteria suggests a mechanism for the diderm-to-monoderm transition. *Nat Microbiol* **7**, 411–422 (2022).

47. V. Sojo, B. Herschy, A. Whicher, E. Camprubí, N. Lane, The Origin of Life in Alkaline Hydrothermal Vents. *Astrobiology* **16**, 181–197 (2016).

48. C. Rinke, P. Schwientek, A. Sczyrba, N. N. Ivanova, I. J. Anderson, J.-F. Cheng, A. Darling, S. Malfatti, B. K. Swan, E. A. Gies, J. A. Dodsworth, B. P. Hedlund, G. Tsiamis, S. M. Sievert, W.-T. Liu, J. A. Eisen, S. J. Hallam, N. C. Kyrpides, R. Stepanauskas, E. M. Rubin, P. Hugenholtz, T. Woyke, Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).

49. J. H. Saw, A. Spang, K. Zaremba-Niedzwiedzka, L. Juzokaite, J. A. Dodsworth, S. K. Murugapiran, D. R. Colman, C. Takacs-Vesbach, B. P. Hedlund, L. Guy, T. J. G. Ettema, Exploring microbial dark matter to resolve the deep archaeal ancestry of eukaryotes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140328 (2015).

50. C. J. Castelle, K. C. Wrighton, B. C. Thomas, L. A. Hug, C. T. Brown, M. J. Wilkins, K. R. Frischkorn, S. G. Tringe, A. Singh, L. M. Markillie, R. C. Taylor, K. H. Williams, J. F. Banfield, Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr. Biol.* **25**, 690–701 (2015).

51. T. A. Williams, G. J. Szöllősi, A. Spang, P. G. Foster, S. E. Heaps, B. Boussau, T. J. G. Ettema, T. M. Embley, Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E4602–E4611 (2017).

52. M. Aouad, N. Taib, A. Oudart, M. Lecocq, M. Gouy, C. Brochier-Armanet, Extreme halophilic archaea derive from two distinct methanogen Class II lineages. *Mol. Phylogenet. Evol.* **127**, 46–54 (2018).

53. Y. Feng, U. Neri, S. Gosselin, A. S. Louyakis, R. T. Papke, U. Gophna, J. P. Gogarten, The Evolutionary Origins of Extreme Halophilic Archaeal Lineages. *Genome Biol. Evol.* **13** (2021).

54. I. A. Berg, Ecological aspects of the distribution of different autotrophic CO2 fixation pathways. *Appl. Environ. Microbiol.* **77**, 1925–1936 (2011).

55. G. Fuchs, Alternative pathways of carbon dioxide fixation: insights into the early evolution of life? *Annu. Rev. Microbiol.* **65**, 631–658 (2011).

56. A. J. Probst, T. Weinmaier, K. Raymann, A. Perras, J. B. Emerson, T. Rattei, G. Wanner, A. Klingl, I. A. Berg, M. Yoshinaga, B. Viehweger, K.-U. Hinrichs, B. C. Thomas, S. Meck, A. K. Auerbach, M. Heise, A. Schintlmeister, M. Schmid, M. Wagner, S. Gribaldo, J. F. Banfield, C. Moissl-Eichinger, Biology of a widespread uncultivated archaeon that contributes to carbon fixation in the subsurface. *Nat. Commun.* **5**, 5497 (2014).

57. P. S. Adam, G. Borrel, S. Gribaldo, Evolutionary history of carbon monoxide dehydrogenase/acetyl-CoA synthase, one of the oldest enzymatic complexes. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E1166–E1173 (2018).

58. I. H. Zhang, B. Borer, R. Zhao, S. Wilbert, D. K. Newman, A. R. Babbin, Uncultivated DPANN archaea are ubiquitous inhabitants of global oxygen deficient zones with diverse metabolic potential. *bioRxiv*, doi: 10.1101/2023.10.30.564641 (2023).

59. E. Biegel, S. Schmidt, J. M. González, V. Müller, Biochemistry, evolution and physiological function of the Rnf complex, a novel ion-motive electron transport complex in prokaryotes. *Cell. Mol. Life Sci.* **68**, 613–634 (2011).

60. L. Westphal, A. Wiechmann, J. Baker, N. P. Minton, V. Müller, The Rnf Complex Is an Energy-Coupled Transhydrogenase Essential To Reversibly Link Cellular NADH and Ferredoxin Pools in the Acetogen Acetobacterium woodii. *J. Bacteriol.* **200** (2018).

61. K. Katoh, K. Misawa, K.-I. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).

62. A. Criscuolo, S. Gribaldo, BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).

63. D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

64. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, R. Lanfear, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

65. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

66. M. Krupovic, V. V. Dolja, E. V. Koonin, The LUCA and its complex virome. *Nat. Rev. Microbiol.* **18**, 661–670 (2020).

67. M. Krupovic, V. V. Dolja, E. V. Koonin, Origin of viruses: primordial replicators recruiting capsids from hosts. *Nat. Rev. Microbiol.* **17**, 449–458 (2019).

APPENDIX A

The nature of the last universal common ancestor and its impact on the early Earth system

Edmund R. R. Moody, Sandra Álvarez-Carretero, Tara A. Mahendrarajah, James W. Clark, Holly C. Betts, Nina Dombrowski, Lénárd L. Szánthó, Richard A. Boyle, Stuart Daines, Xi Chen, Nick Lane, Ziheng Yang, Graham A. Shields, Gergely J. Szöllősi, Anja Spang, Davide Pisani, Tom A. Williams, Timothy M. Lenton & Philip C. J. Donoghue

Nature Ecology and Evolution (2024)

ABSTRACT

The nature of the last universal common ancestor (LUCA), its age and its impact on the Earth system have been the subject of vigorous debate across diverse disciplines, often based on disparate data and methods. Age estimates for LUCA are usually based on the fossil record, varying with every reinterpretation. The nature of LUCA's metabolism has proven equally contentious, with some attributing all core metabolisms to LUCA, whereas others reconstruct a simpler life form dependent on geochemistry. Here we infer that LUCA lived ~4.2 Ga (4.09–4.33 Ga) through divergence time analysis of pre-LUCA gene duplicates, calibrated using microbial fossils and isotope records under a new cross-bracing implementation. Phylogenetic reconciliation suggests that LUCA had a genome of at least 2.5 Mb (2.49–2.99 Mb), encoding around 2,600 proteins, comparable to modern prokaryotes. Our results suggest LUCA was a prokaryote-grade anaerobic acetogen that possessed an early immune system. Although LUCA is sometimes perceived as living in isolation, we infer LUCA to have been part of an established ecological system. The metabolism of LUCA would have provided a niche for other microbial community members and hydrogen recycling by atmospheric photochemistry could have supported a modestly productive early ecosystem.

MAIN

The common ancestry of all extant cellular life is evidenced by the universal genetic code, machinery for protein synthesis, shared chirality of the almost-universal set of 20 amino acids and use of ATP as a common energy currency (*1*). The last universal common ancestor (LUCA) is the node on the tree of life from which the fundamental prokaryotic domains (Archaea and Bacteria) diverge. As such, our understanding of LUCA impacts our understanding of the early evolution of life on Earth. Was LUCA a simple or complex organism? What kind of environment did it inhabit and when? Previous estimates of LUCA are in conflict either due to conceptual disagreement about what LUCA is (*2*) or as a result of different methodological approaches and data (*3*–*9*). Published analyses differ in their inferences of LUCA's genome, from conservative estimates of 80 orthologous proteins (*10*) up to 1,529 different potential gene families (*4*). Interpretations range from little beyond an information-processing and metabolic core (*6*) through to a prokaryote-grade organism with much of the gene repertoire of modern Archaea and Bacteria (*8*), recently reviewed in ref. (*7*). Here we use molecular clock methodology, horizontal gene-transfer-aware phylogenetic reconciliation and existing biogeochemical models to address questions about LUCA's age, gene content, metabolism and impact on the early Earth system.

ESTIMATING THE AGE OF LUCA

Life's evolutionary timescale is typically calibrated to the oldest fossil occurrences. However, the veracity of fossil discoveries from the early Archaean period has been contested (*11*, *12*). Relaxed Bayesian node-calibrated molecular clock approaches provide a means of integrating the sparse fossil and geochemical record of early life with the information provided by molecular data; however, constraining LUCA's age is challenging due to limited prokaryote fossil calibrations and the uncertainty in their placement on the phylogeny. Molecular clock estimates of LUCA (*13–15*) have relied on conserved universal single-copy marker genes within phylogenies for which LUCA represented the root. Dating the root of a tree is difficult because errors propagate from the tips to the root of the dated phylogeny and information is not available to estimate the rate of evolution for the branch incident on the root node. Therefore, we analysed genes that duplicated before LUCA with two (or more) copies in LUCA's genome (*16*). The root in these gene trees represents this duplication preceding LUCA, whereas LUCA is represented by two descendant nodes. Use of these universal paralogues also has the advantage that the same calibrations can be applied at least twice. After duplication, the same species divergences are represented on both sides of the gene tree (*17*, *18*) and thus can be assumed to have the same age. This considerably reduces the uncertainty when genetic distance (branch length) is resolved into absolute time and rate. When a shared node is assigned a fossil calibration, such cross-bracing also serves to double the number of calibrations on the phylogeny, improving divergence time estimates. We calibrated our molecular clock analyses using 13 calibrations (see 'Fossil calibrations' in Supplementary Information). The calibration on the root of the tree of life is of particular importance. Some previous studies have placed a younger maximum constraint on the age of LUCA based on the assumption that life could not have survived Late Heavy Bombardment (LHB) (~3.7–3.9 billion years ago (Ga)) (*19*). However, the LHB hypothesis is extrapolated and scaled from the Moon's impact record, the interpretation of which has been questioned in terms of the intensity, duration and even the veracity of an LHB episode (*20–23*). Thus, the LHB hypothesis should not be considered a credible maximum constraint on the age of LUCA. We used soft-uniform bounds, with the maximum-age bound based on the time of the Moon-forming impact (4,510 million years ago (Ma) ± 10 Myr), which would have effectively sterilized Earth's precursors, Tellus and Theia (*13*). Our minimum bound on the age of LUCA is based on low δ^{98}Mo isotope values indicative of Mn oxidation compatible with oxygenic photosynthesis and, therefore, total-group Oxyphotobacteria in the Mozaan Group, Pongola Supergroup, South Africa (*24*, *25*), dated minimally to 2,954 Ma ± 9 Myr (ref. (*26*)).

Our estimates for the age of LUCA are inferred with a concatenated and a partitioned dataset, both consisting of five pre-LUCA paralogues: catalytic and non-catalytic subunits from ATP synthases, elongation factor Tu and G, signal recognition protein and signal recognition particle receptor, tyrosyl-tRNA and tryptophanyl-tRNA synthetases, and leucyl- and valyl-tRNA synthetases (*27*). Marginal densities (commonly referred to as effective priors) fall within calibration densities (that is, user-specified priors) when topologically adjacent

calibrations do not overlap temporally, but may differ when they overlap, to ensure the relative age relationships between ancestor-descendant nodes. We consider the marginal densities a reasonable interpretation of the calibration evidence given the phylogeny; we are not attempting to test the hypothesis that the fossil record is an accurate temporal archive of evolutionary history because it is not (*28*). The duplicated LUCA node age estimates we obtained under the autocorrelated rates (geometric Brownian motion (GBM)) (*29*, *30*) and independent-rates log-normal (ILN) (*31*, *32*) relaxed-clock models with our partitioned dataset (GBM, 4.18–4.33 Ga; ILN, 4.09–4.32 Ga; Fig. 1) fall within our composite age estimate for LUCA ranging from 3.94 Ga to 4.52 Ga, comparable to previous studies (*13*, *18*, *33*). Dating analyses based on single genes, or concatenations that excluded each gene in turn, returned compatible timescales (Extended Data Figs. 1 and 2 and 'Additional methods' in Methods).

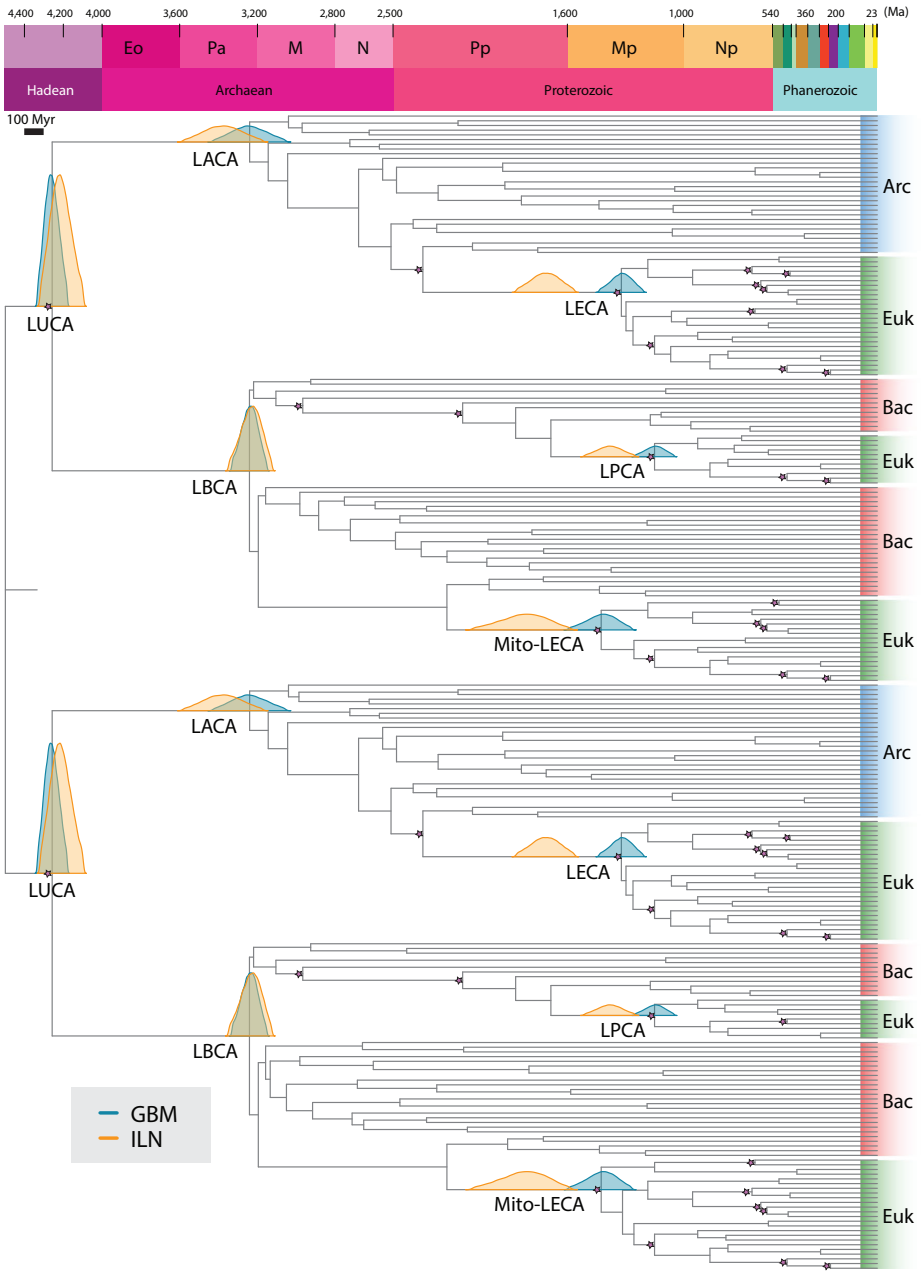


Fig. 1 | Timetree inferred under a Bayesian node-dating approach with cross-bracing using a partitioned dataset of five pre-LUCA paralogues. Our results suggest that LUCA lived around 4.2 Ga, with a 95% confidence interval spanning 4.09–4.33 Ga under the ILN relaxed-clock model (orange) and 4.18–4.33 Ga under the GBM relaxed-clock model (teal). Under a cross-bracing approach, nodes corresponding to the same species divergences (that is, mirrored nodes) have the same posterior time densities. This figure shows the corresponding posterior time densities of the mirrored nodes for the last universal, archaeal, bacterial and eukaryotic common ancestors (LUCA, LACA, LBCA and LECA,

respectively); the last common ancestor of the mitochondrial lineage (Mito-LECA); and the last plastid-bearing common ancestor (LPCA). Purple stars indicate nodes calibrated with fossils. Arc, Archaea; Bac, Bacteria; Euk, Eukarya.

LUCA'S PHYSIOLOGY

To estimate the physiology of LUCA, we first inferred an updated microbial phylogeny from 57 phylogenetic marker genes (see 'Universal marker genes' in Methods) on 700 genomes, comprising 350 Archaea and 350 Bacteria (*15*). This tree was in good agreement with recent phylogenies of the archaeal and bacterial domains of life (*34*, *35*). For example, the TACK (*36*) and Asgard clades of Archaea (*37–39*) and Gracilicutes within Bacteria (*40*, *41*) were recovered as monophyletic. However, the analysis was equivocal as to the phylogenetic placement of the Patescibacteria (CPR) (*42*) and DPANN (*43*), which are two small-genome lineages that have been difficult to place in trees. Approximately unbiased (*44*) tests could not distinguish the placement of these clades, neither at the root of their respective domains nor in derived positions, with CPR sister to Chloroflexota (as reported recently in refs. (*35*, *41*, *45*)) and DPANN sister to Euryarchaeota. To account for this phylogenetic uncertainty, we performed LUCA reconstructions on two trees: our maximum likelihood (ML) tree (topology 1; Extended Data Fig. 3) and a tree in which CPR were placed as the sister of Chloroflexota, with DPANN sister to all other Archaea (topology 2; Extended Data Fig. 4). In both cases, the gene families mapped to LUCA were very similar (correlation of LUCA presence probabilities (PP), $r = 0.6720275$, $P < 2.2 \times 10^{-16}$). We discuss the results on the tree with topology 2 and discuss the residual differences in Supplementary Information, 'Topology 1' (Supplementary Data 1).

We used the probabilistic gene- and species-tree reconciliation algorithm ALE (*46*) to infer the evolution of gene family trees for each sampled entry in the KEGG Orthology (KO) database (*47*) on our species tree. ALE infers the history of gene duplications, transfers and losses based on a comparison between a distribution of bootstrapped gene trees and the reference species tree, allowing us to estimate the probability that the gene family was present at a node in the tree (*35*, *48*, *49*). This reconciliation approach has several advantages for drawing inferences about LUCA. Most gene families have experienced gene transfer since the time of LUCA (*50*, *51*) and so explicitly modelling transfers enables us to include many more gene families in the analysis than has been possible using previous approaches. As the analysis is probabilistic, we can also account for uncertainty in gene family origins and evolutionary history by averaging over different scenarios using the reconciliation model. Using this approach, we estimated the probability that each KEGG gene family (KO) was present in LUCA and then used the resulting probabilities to construct a hypothetical model of LUCA's gene content, metabolic potential (Fig. 2) and environmental context (Fig. 3). Using the KEGG annotation is beneficial because it allows us to connect our inferences to curated functional annotations; however, it has the drawback that some widespread gene families that were likely present in LUCA are divided into multiple KO families that individually appear to be restricted to particular taxonomic groups and inferred to have arisen later. To account for this limitation, we also performed an

analysis of COG (Clusters of Orthologous Genes) (*52*) gene families, which correspond to more coarse-grained functional annotations (Supplementary Data 2).

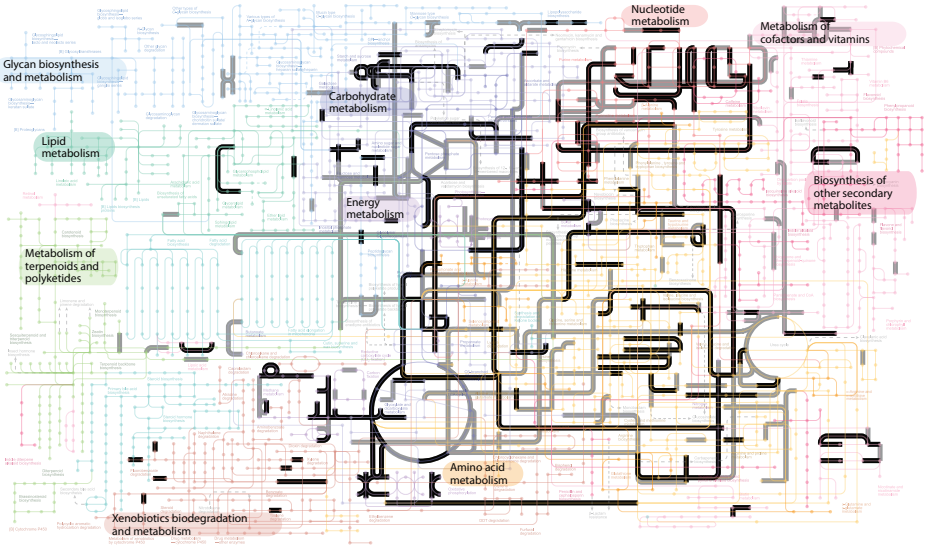


Fig. 2 | **Probabilistic estimates of metabolic networks from modern life that were present in LUCA.** In black: enzymes and metabolic pathways inferred to be present in LUCA with at least PP = 0.75, with sampling in both prokaryotic domains. In grey: those inferred in our least-stringent threshold of PP = 0.50. The analysis supports the presence of a complete WLP and an almost complete TCA cycle across multiple confidence thresholds. Metabolic maps derived from KEGG *(47)* database through iPath *(109).* GPI, glycosylphosphatidylinositol; DDT, 1,1,1-trichloro-2,2-bis(p-chlorophenyl)ethane.

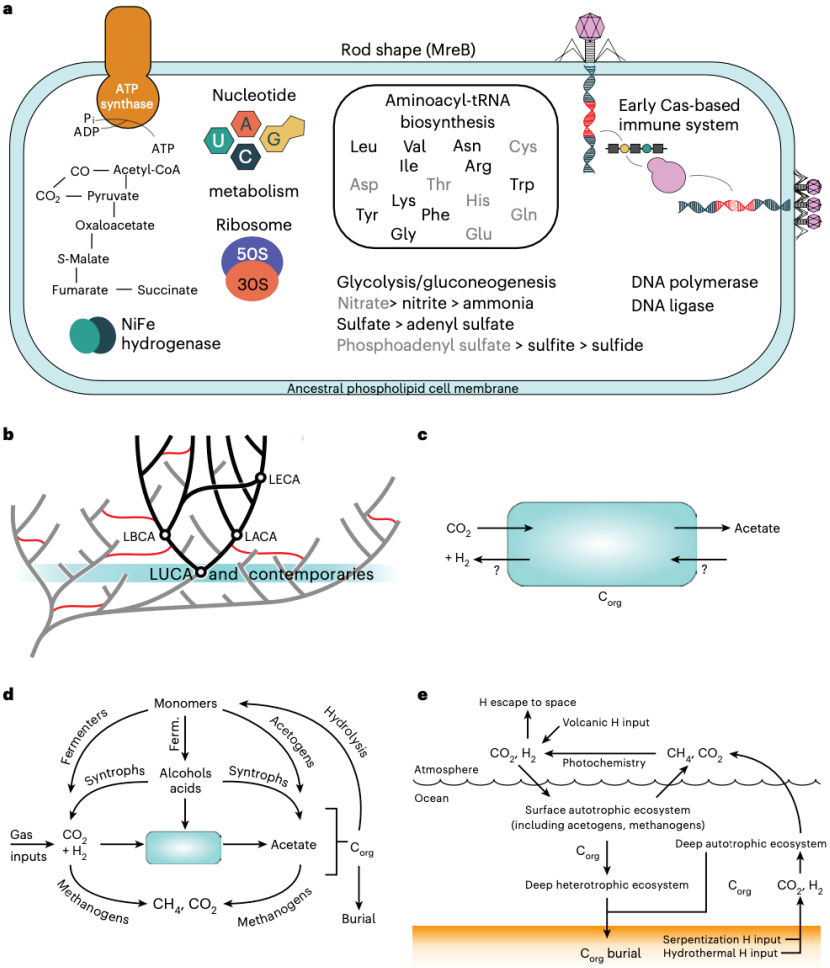


Fig. 3 | A reconstruction of LUCA, within its evolutionary and ecological context. a, A representation of LUCA based on our ancestral gene content reconstruction. Gene names in black have been inferred to be present in LUCA under the most-stringent threshold (PP = 0.75, sampled in both domains); those in grey are present at the least-stringent threshold (PP = 0.50, without a requirement for presence in both domains). **b**, LUCA in the context of the tree of life. Branches on the tree of life that have left sampled descendants today are coloured black, those that have left no sampled descendants are in grey. As the common ancestor of extant cellular life, LUCA is the oldest node that can be reconstructed using phylogenetic methods. It would have shared the early Earth with other lineages (highlighted in teal) that have left no descendants among sampled cellular life today. However, these lineages may have left a trace in modern organisms by transferring genes into the sampled tree of life (red lines) before their extinction. **c**, LUCA's chemoautotrophic metabolism probably relied on gas exchange with the immediate environment to achieve organic carbon (C_{org}) fixation via acetogenesis and it may also have run the metabolism in reverse. **d**, LUCA within the context of an early ecosystem. The CO_2 and H_2 that fuelled LUCA's plausibly acetogenic metabolism could have come from both geochemical and biotic inputs. The organic matter and acetate that LUCA produced could have created a niche for other metabolisms, including ones that recycled CO_2 and H_2 (as in modern sediments). **e**, LUCA in an Earth system context. Acetogenic LUCA could have been a key part of both surface and deep (chemo)autotrophic ecosystems, powered by H_2. If methanogens were also present, hydrogen would be released as CH_4 to the atmosphere, converted to H_2 by photochemistry and thus recycled back to the surface ecosystem, boosting its productivity. Ferm., fermentation.

GENOME SIZE AND CELLULAR FEATURES

By using modern prokaryotic genomes as training data, we used a predictive model to estimate the genome size and the number of protein families encoded by LUCA based on the relationship between the number of KEGG gene families and the total number of proteins encoded by modern prokaryote genomes (Extended Data Figs. 5 and 6). On the basis of the PPs for KEGG KO gene families, we identified a conservative subset of 399 KOs that were likely to be present in LUCA, with PPs ≥0.75, and found in both Archaea and Bacteria (Supplementary Data 1); these families form the basis of our metabolic reconstruction. However, by integrating over the inferred PPs of all KO gene families, including those with low probabilities, we also estimate LUCA's genome size. Our predictive model estimates a genome size of 2.75 Mb (2.49–2.99 Mb) encoding 2,657 (2,451–2,855) proteins (Methods). Although we can estimate the number of genes in LUCA's genome, it is more difficult to identify the specific gene families that might have already been present in LUCA based on the genomes of modern Archaea and Bacteria. It is likely that the modern version of the pathways would be considered incomplete based on LUCA's gene content through subsequent evolutionary changes. We should therefore expect reconstructions of metabolic pathways to be incomplete due to this phylogenetic noise and other limitations of the analysis pipeline. For example, when looking at genes and pathways that can uncontroversially be mapped to LUCA, such as the ribosome and aminoacyl-tRNA synthetases for implementing the genetic code, we find that we map many (but not all) of the key components to LUCA (see 'Notes' in Supplementary Information). We interpret this to mean that our reconstruction is probably incomplete but our interpretation of LUCA's metabolism relies on our inference of pathways, not individual genes.

The inferred gene content of LUCA suggests it was an anaerobe as we do not find support for the presence of terminal oxidases (Supplementary Data 1). Instead we identified almost all genes encoding proteins of the archaeal (and most of the bacterial) versions of the Wood–Ljungdahl pathway (WLP) (PP > 0.7), indicating that LUCA had the potential for acetogenic growth and/or carbon fixation (*53–55*) (Supplementary Data 3). LUCA encoded some NiFe hydrogenase subunits (K06281, PP = 0.90; K14126, PP = 0.92), which may have enabled growth on hydrogen (see 'Notes' in Supplementary Information). Complexes involved in methanogenesis such as methyl-coenzyme M reductase and tetrahydromethanopterin S-methyltransferase were inferred to be absent, suggesting that LUCA was unlikely to function as a modern methanogen. We found strong support for some components of the TCA cycle (including subunits of oxoglutarate/2-oxoacid ferredoxin oxidoreductase (K00175 and K00176), succinate dehydrogenase (K00239) and homocitrate synthase (K02594)), although some steps are missing. LUCA was probably capable of gluconeogenesis/glycolysis in that we find support for most subunits of enzymes involved in these pathways (Supplementary Data 1 and 3). Considering the presence of the WLP, this may indicate that LUCA had the ability to grow organoheterotrophically and potentially also autotrophically. Gluconeogenesis would have been important in linking carbon fixation to nucleotide biosynthesis via the pentose phosphate pathway, most enzymes of which seem to be present in LUCA (see 'Notes' in Supplementary

Information). We found no evidence that LUCA was photosynthetic, with low PPs for almost all components of oxygenic and anoxygenic photosystems (Supplementary Data 3).

We find strong support for the presence of ATP synthase, specifically, the A (K02117, PP = 0.98) and B (K02118, PP = 0.94) subunit components of the hydrophilic V/A1 subunit, and the I (subunit a, K02123, PP = 0.99) and K (subunit c, K02124, PP = 0.82) subunits of the transmembrane V/A0 subunit. In addition, if we relax the sampling threshold, we also infer the presence of the F1-type β-subunit (K02112, PP = 0.94). This is consistent with many previous studies that have mapped ATP synthase subunits to LUCA (*6*, *17*, *18*, *56*, *57*).

We obtain moderate support for the presence of pathways for assimilatory nitrate (ferredoxin-nitrate reductase, K00367, PP = 0.69; ferredoxin-nitrite reductase, K00367, PP = 0.53) and sulfate reduction (sulfate adenylyltransferase, K00957, PP = 0.80, and K00958, PP = 0.73; sulfite reductase, K00392, PP = 0.82; phosphoadenosine phosphosulfate reductase, K00390, PP = 0.56), probably to fuel amino acid biosynthesis, for which we inferred the presence of 37 partially complete pathways.

We found support for the presence of 19 class 1 CRISPR–Cas effector protein families in the genome of LUCA, including types I and III (cas3, K07012, PP = 0.80, and K07475, PP = 0.74; cas10, K07016, PP = 0.96, and K19076, PP = 0.67; and cas7, K07061, PP = 0.90, K09002, PP = 0.84, K19075, PP = 0.97, K19115, PP = 0.98, and K19140, PP = 0.80). The absence of Cas1 and Cas2 may suggest LUCA encoded an early Cas system with the means to deliver an RNA-based immune response by cutting (Cas6/Cas3) and binding (CSM/Cas10) RNA, but lacking the full immune-system-site CRISPR. This supports the idea that the effector stage of CRISPR–Cas immunity evolved from RNA sensing for signal transduction, based on the similarities in RNA binding modules of the proteins (*58*). This is consistent with the idea that cellular life was already involved in an arms race with viruses at the time of LUCA (*59*, *60*). Our results indicate that an early Cas system was an ancestral immune system of extant cellular life.

Altogether, our metabolic reconstructions suggest that LUCA was a relatively complex organism, similar to extant Archaea and Bacteria (*6*, *7*). On the basis of ancient duplications of the *Sec* and ATP synthase genes before LUCA, along with high PPs for key components of those systems, membrane-bound ATP synthase subunits, genes involved in peptidoglycan synthesis (*mraY*, K01000; *murC*, K01924) and the cytoskeletal actin-like protein, MreB (K03569) (Supplementary Data 3), it is highly likely that LUCA possessed the core cellular apparatus of modern prokaryotic life. This might include the basic constituents of a phospholipid membrane, although our analysis did not conclusively establish its composition. In particular, we recovered the following enzymes involved in the synthesis of ether and ester lipids, (alkyldihydroxyacetonephosphate synthase, glycerol 3-phosphate and glycerol 1-phosphate) and components of the mevalonate pathway (mevalonate 5-phosphate dehydratase (PP = 0.84), hydroxymethylglutaryl-CoA reductase (PP = 0.52), mevalonate kinase (PP = 0.51) and hydroxymethylglutaryl-CoA synthase (PP = 0.51)).

Compared with previous estimates of LUCA's gene content, we find 81 overlapping COG gene families with the consensus dataset of ref. (*7*) and 69 overlapping KOs with the dataset of ref. (*6*). Key points of agreement between previous studies include the presence of signal recognition particle protein, ffh (COG0541, K03106) (*7*) used in the targeting and delivery of proteins for the plasma membrane, a high number of aminoacyl-tRNA synthetases for amino acid synthesis and glycolysis/gluconeogenesis enzymes.

Ref. (*6*) inferred LUCA to be a thermophilic anaerobic autotroph using the WLP for carbon fixation based on the presence of a single enzyme (CODH), and similarly suggested that LUCA was capable of nitrogen fixation using a nitrogenase. Our reconstruction agrees with ref. (*6*) that LUCA was an anaerobic autotroph using the WLP for carbon fixation, but we infer the presence of a much more complete WLP than that previously obtained. We did not find strong evidence for nitrogenase or nitrogen fixation, and the reconstruction was not definitive with respect to the optimal growth environment of LUCA.

We used a probabilistic approach to reconstruct LUCA—that is, we estimated the probability with which each gene family was present in LUCA based on a model of how gene families evolve along an overarching species tree. This approach differs from analyses of phylogenetic presence–absence profiles (*3*, *4*, *9*) or those that used filtering criteria (such as broadly distributed or highly vertically evolving families) to define a high-confidence subset of modern genes that might have been present in LUCA. Our reconstruction maps many more genes to LUCA—albeit each with lower probability—than previous analyses (*8*) and yields an estimate of LUCA's genome size that is within the range of modern prokaryotes. The result is an incomplete picture of a cellular organism that was prokaryote grade rather than progenotic (*2*) and that, similarly to prokaryotes today, probably existed as part of an ecosystem. As the common ancestor of sampled, extant prokaryotic life, LUCA is the oldest node on the species tree that we can reconstruct via phylogenomics but, as Fig. 3 illustrates, it was already the product of a highly innovative period in evolutionary history during which most of the core components of cells were established. By definition, we cannot reconstruct LUCA's contemporaries using phylogenomics but we can propose hypotheses about their physiologies based on the reconstructed LUCA whose features immediately suggest the potential for interactions with other prokaryotic metabolisms.

LUCA'S ENVIRONMENT, ECOSYSTEM AND EARTH SYSTEM CONTEXT

The inference that LUCA used the WLP helps constrain the environment and ecology in which it could have lived. Modern acetogens can grow autotrophically on H_2 (and CO_2) or heterotrophically on a wide range of alternative electron donors including alcohols, sugars and carboxylic acids (*55*). This metabolic flexibility is key to their modern ecological success. Acetogenesis, whether autotrophic or heterotrophic, has a low energy yield and growth efficiency (although use of the reductive acetyl-CoA pathway for both energy production

and biosynthesis reduces the energy cost of biosynthesis). This would be consistent with an energy-limited early biosphere (*61*).

If LUCA functioned as an organoheterotrophic acetogen, it was necessarily part of an ecosystem containing autotrophs providing a source of organic compounds (because the abiotic source flux of organic molecules was minimal on the early Earth). Alternatively, if LUCA functioned as a chemoautotrophic acetogen it could (in principle) have lived independently off an abiotic source of H_2 (and CO_2). However, it is implausible that LUCA would have existed in isolation as the by-products of its chemoautotrophic metabolism would have created a niche for a consortium of other metabolisms (as in modern sediments) (Fig. 3d). This would include the potential for LUCA itself to grow as an organoheterotroph.

A chemoautotrophic acetogenic LUCA could have occupied two major potential habitats (Fig. 3e): the first is the deep ocean where hydrothermal vents and serpentinization of sea-floor provided a source of H_2 (ref. (*62*)). Consistent with this, we find support for the presence of reverse gyrase (PP = 0.97), a hallmark enzyme of hyperthermophilic prokaryotes (*6*, *63–65*), which would not be expected if early life existed at the ocean surface (although the evolution of reverse gyrase is complex (*63*); see 'Reverse gyrase' in Supplementary Information). The second habitat is the ocean surface where the atmosphere would have provided a source of H_2 derived from volcanoes and metamorphism. Indeed, we detected the presence of spore photoproduct lyase (COG1533, K03716, PP = 0.88) that in extant organisms repairs methylene-bridged thymine dimers occurring in spore DNA as a result of damage induced through ultraviolet (UV) radiation (*66*, *67*). However, this gene family also occurs in modern taxa that neither form endospores nor dwell in environments where they are likely to accrue UV damage to their DNA and so is not an exclusive hallmark of environments exposed to UV. Previous studies often favoured a deep-ocean environment for LUCA as early life would have been better protected there from an episode of LHB. However, if the LHB was less intense than initially proposed (*20*, *22*), or just a sampling artefact (*21*), these arguments weaken. Another possibility may be that LUCA inhabited a shallow hydrothermal vent or a hot spring.

Hydrogen fluxes in these ecosystems could have been several times higher on the early Earth (with its greater internal heat source) than today. Volcanism today produces ~1×10^{12} mol H_2 yr^{-1} and serpentinization produces ~0.4×10^{12} mol H_2 yr^{-1}. With the present H_2 flux and the known scaling of the H_2 escape rate to space, an abiotic atmospheric concentration of H_2 of ~150 ppmv is predicted (*68*). Chemoautotrophic acetogens would have locally drawn down the concentration of H_2 (in either surface or deep niche) but their low growth efficiency would ensure H_2 (and CO_2) remained available. This and the organic matter and acetate produced would have created niches for other metabolisms, including methanogenesis (Fig. 3d).

On the basis of thermodynamic considerations, CH_4 and CO_2 are expected to be the eventual metabolic end products of the resulting ecosystem, with a small fraction of the initial hydrogen consumption buried as organic matter. The resulting flux of CH_4 to the atmosphere would fuel

photochemical H_2 regeneration and associated productivity in the surface ocean (Fig. 3e). Existing models suggest the resulting global H_2 recycling system is highly effective, such that the supply flux of H_2 to the surface could have exceeded the volcanic input of H_2 to the atmosphere by at least an order of magnitude, in turn implying that the productivity of such a biosphere was boosted by a comparable factor (*69*). Photochemical recycling to CO would also have supported a surface niche for organisms consuming CO (ref. (*69*)).

In deep-ocean habitats, there could be some localized recycling of electrons (Fig. 3d) but a quantitative loss of highly insoluble H_2 and CH_4 to the atmosphere and minimal return after photochemical conversion of CH_4 to H_2 means global recycling to depth would be minimal (Fig. 3e). Hence the surface environment for LUCA could have become dominant (albeit recycling of the resulting organic matter could be spread through ocean depth; 'Deep heterotrophic ecosystem' in Fig. 3e). The global net primary productivity of an early chemoautotrophic biosphere including acetogenic LUCA and methanogens could have been of order ~1×10^{12} to 7×10^{12} mol C yr^{-1} (~3 orders of magnitude less than today) (*69*).

The nutrient supply (for example, N) required to support such a biosphere would need to balance that lost in the burial flux of organic matter. Earth surface redox balance dictates that hydrogen loss to space and burial of electrons/hydrogen must together balance input of electrons/hydrogen. Considering contemporary H_2 inputs, and the above estimate of net primary productivity, this suggests a maximum burial flux in the order of ~10^{12} mol C yr^{-1}, which, with contemporary stoichiometry (C:N ratio of ~7) could demand >10^{11} mol N yr^{-1}. Lightning would have provided a source of nitrite and nitrate (*70*), consistent with LUCA's inferred pathways of nitrite and (possibly) nitrate reduction. However, it would only have been of the order 3×10^9 mol N yr^{-1} (ref. (*71*)). Instead, in a global hydrogen-recycling system, HCN from photochemistry higher in the atmosphere, deposited and hydrolysed to ammonia in water, would have increased available nitrogen supply by orders of magnitude toward ~3×10^{12} mol N yr^{-1} (refs. (*71*, *72*)). This HCN pathway is consistent with the anomalously light nitrogen isotopic composition of the earliest plausible biogenic matter of 3.8–3.7 Ga (ref. (*73*)), although that considerably postdates our inferred age of LUCA. These considerations suggest that the proposed LUCA biosphere (Fig. 3e) would have been energy or hydrogen limited not nitrogen limited.

CONCLUSIONS

By treating gene presence probabilistically, our reconstruction maps many more genes (2,657) to LUCA than previous analyses and results in an estimate of LUCA's genome size (2.75 Mb) that is within the range of modern prokaryotes. The result is a picture of a cellular organism that was prokaryote grade rather than progenotic (*2*) and that probably existed as a component of an ecosystem, using the WLP for acetogenic growth and carbon fixation. We cannot use phylogenetics to reconstruct other members of this early ecosystem but we can infer their

physiologies based on the metabolic inputs and outputs of LUCA. How evolution proceeded from the origin of life to early communities at the time of LUCA remains an open question, but the inferred age of LUCA (~4.2 Ga) compared with the origin of the Earth and Moon suggests that the process required a surprisingly short interval of geologic time.

METHODS

UNIVERSAL MARKER GENES

A list of 298 markers were identified by creating a non-redundant list of markers used in previous studies on archaeal and bacterial phylogenies (*10*, *35*, *38*, *74–79*). These markers were mapped to the corresponding COG, arCOG and TIGRFAM profile to identify which profile is best suited to extract proteins from taxa of interest. To evaluate whether the markers cover all archaeal and bacterial diversity, proteins from a set of 574 archaeal and 3,020 bacterial genomes were searched against the COG, arCOG and TIGRFAM databases using hmmsearch (v.3.1b2; settings, hmmsearch–tblout output–domtblout–notextw) (*52*, *80–82*). Only hits with an e-value less than or equal to 1×10^{-5} were investigated further and for each protein the best hit was determined based on the e-value (expect value) and bit-score. Results from all database searches were merged based on the protein identifiers and the table was subsetted to only include hits against the 298 markers of interest. On the basis of this table we calculated whether the markers occurred in Archaea, Bacteria or both Archaea and Bacteria. Markers were only included if they were present in at least 50% of taxa and contained less than 10% of duplications, leaving a set of 265 markers. Sequences for each marker were aligned using MAFFT L-INS-i v.7.407 (ref. (*83*)) for markers with less than 1,000 sequences or MAFFT (*84*) for those with more than 1,000 sequences (setting, –reorder) (*84*) and sequences were trimmed using BMGE (*85*) set for amino acids, a BLOcks SUbstitution Matrix 30 similarity matrix, with a entropy score of 0.5 (v.1.12; settings, -t AA -m BLOSUM30 -h 0.5). Single gene trees were generated with IQ-TREE 2 (ref. (*86*)), using the LG substitution matrix, with ten-profile mixture models, four CPUs, with 1,000 ultrafast bootstraps optimized by nearest neighbour interchange written to a file retaining branch lengths (v.2.1.2; settings, -m LG+C10+F+R -nt 4 -wbtl -bb 1,000 -bnni). These single gene trees were investigated for archaeal and bacterial monophyly and the presence of paralogues. Markers that failed these tests were not included in further analyses, leaving a set of 59 markers (3 arCOGs, 46 COGs and 10 TIGRFAMs) suited for phylogenies containing both Archaea and Bacteria (Supplementary Data 4).

MARKER GENE SEQUENCE SELECTION

To limit selecting distant paralogues and false positives, we used a bidirectional or reciprocal approach to identify the sequences corresponding to the 59 single-copy markers. In the first inspection (query 1), the 350 archaeal and 350 bacterial reference genomes were queried against all arCOG HMM (hidden Markov model) profiles (All_Arcogs_2018.hmm), all COG HMM profiles (NCBI_COGs_Oct2020.hmm) and all TIGRFAM HMM profiles (TIGRFAMs_15.0_HMM.LIB) using a custom script built on hmmsearch: hmmsearchTable <genomes.faa><database.

hmm>-E 1×10^{-5} >HMMscan_Output_e5 (HMMER v.3.3.2) (*87*). HMM profiles corresponding to the 59 single-copy marker genes (Supplementary Data 4) were extracted from each query and the best-hit sequences were identified based on the e-value and bit-score. We used the same custom hmmsearchTable script and conditions (see above) in the second inspection (query 2) to query the best-hit sequences identified above against the full COG HMM database (NCBI_COGs_Oct2020.hmm). Results were parsed and the COG family assigned in query 2 was compared with the COG family assigned to sequences based on the marker gene identity (Supplementary Data 4). Sequence hits were validated using the matching COG identifier, resulting in 353 mismatches (that is, COG family in query 1 does not match COG family in query 2) that were removed from the working set of marker gene sequences. These sequences were aligned using MAFFT L-INS-i (*83*) and then trimmed using BMGE (*85*) with a BLOSUM30 matrix. Individual gene trees were inferred under ML using IQ-TREE 2 (ref. (*86*)) with model fitting, including both the default homologous substitution models and the following complex heterogeneous substitution models (LG substitution matrices with 10–60-profile mixture models, with empirical base frequencies and a discrete gamma model with four categories accounting for rate heterogeneity across sites): LG+C60+F+G, LG+C50+F+G, LG+C40+F+G, LG+C30+F+G, LG+C20+F+G and LG+C10+F+G, with 10,000 ultrafast bootstraps and 10 independent runs to avoid local optima. These 59 gene trees were manually inspected and curated over multiple rounds. Any horizontal gene transfer events, paralogous genes or sequences that violated domain monophyly were removed and two genes (arCOG01561, *tuf*; COG0442, *ProS*) were dropped at this stage due to the high number of transfer events, resulting in 57 single-copy orthologues for further tree inference.

SPECIES-TREE INFERENCE

These 57 orthologous sequences were concatenated and ML trees were inferred after three independent runs with IQ-TREE 2 (ref. (*86*)) using the same model fitting and bootstrap settings as described above. The tree with the highest log-likelihood of the three runs was chosen as the ML species tree (topology 1). To test the effect of removing the CPR bacteria, we removed all CPR bacteria from the alignment before inferring a species tree (same parameters as above). We also performed approximately unbiased (*44*) tree topology tests (with IQ-TREE 2 (ref. (*86*)), using LG+C20+F+G) when testing the significance of constraining the species-tree topology (ML tree; Supplementary Fig. 1) to have a DPANN clade as sister to all other Archaea (same parameters as above but with a minimally constrained topology with monophyletic Archaea and DPANN sister to other Archaea present in a polytomy (Supplementary Fig. 2)) and testing a constraint of CPR to be sister to Chloroflexi (Supplementary Fig. 3), and a combination of both the DPANN and CPR constraints (topology 2); these were tested against the ML topology, both using the normal 20 amino acid alignments and also with Susko–Roger recoding (*88*).

GENE FAMILIES

For the 700 representative species (*15*), gene family clustering was performed using EGGNOGMAPPER v.2 (ref. (*89*)), with the following parameters: using the DIAMOND (*90*) search, a query cover of 50% and an e-value threshold of 0.0000001. Gene families were collated using

their KEGG (*47*) identifier, resulting in 9,365 gene families. These gene families were then aligned using MAFFT (*84*) v.7.5 with default settings and trimmed using BMGE (*85*) (with the same settings as above). Five independent sets of ML trees were then inferred using IQ-TREE 2 (ref. (*86*)), using LG+F+G, with 1,000 ultrafast bootstrap replicates. We also performed a COG-based clustering analysis in which COGs were assigned based on the modal COG identifier annotated for each KEGG gene family based on the results from EGGNOGMAPPER v.2 (ref. (*89*)). These gene families were aligned, trimmed and one set of gene trees (with 1,000 ultrafast bootstrap replicates) was inferred using the same parameters as described above for the KEGG gene families.

RECONCILIATIONS

The five sets of bootstrap distributions were converted into ALE files, using ALEobserve, and reconciled against topology 1 and topology 2 using ALEml_undated (*91*) with the fraction missing for each genome included (where available). Gene family root origination rates were optimized for each COG functional category as previously described (*35*) and families were categorized into four different groups based on the probability of being present in the LUCA node in the tree. The most-stringent category was that with sampling above 1% in both domains and a PP≥0.75, another category was with PP≥0.75 with no sampling requirement, another with PP≥0.5 with the sampling requirement; the least stringent was PP≥0.5 with no sampling requirement. We used the median probability at the root from across the five runs to avoid potential biases from failed runs in the mean and to account for variation across bootstrap distributions (see Supplementary Fig. 4 for distributions of the inferred ratio of duplications, transfers and losses for all gene families across all tips in the species tree; see Supplementary Data 5 for the inferred duplications, transfers and losses ratios for LUCA, the last bacterial common ancestor and the last archaeal common ancestor).

METABOLIC PATHWAY ANALYSIS

Metabolic pathways for gene families mapped to the LUCA node were inferred using the KEGG (*47*) website GUI and metabolic completeness for individual modules was estimated with Anvi'o (*92*) (anvi-estimate-metabolism), with pathwise completeness.

ADDITIONAL TESTING

We tested for the effects of model complexity on reconciliation by using posterior mean site frequency LG+C20+F+G across three independent runs in comparison with 3 LG+F+G independent runs. We also performed a 10% subsampling of the species trees and gene family alignments across two independent runs for two different subsamples, one with and one without the presence of Asgard archaea. We also tested the likelihood of the gene families under a bacterial root (between Terrabacteria and Gracilicutes) using reconciliations of the gene families under a species-tree topology rooted as such.

FOSSIL CALIBRATIONS

On the basis of well-established geological events and the fossil record, we modelled 13 uniform densities to constrain the maximum and minimum ages of various nodes in our

phylogeny. We constrained the bounds of the uniform densities to be either hard (no tail probability is allowed after the age constraint) or soft (a 2.5% tail probability is allowed after the age constraint) depending on the interpretation of the fossil record (Supplementary Information). Nodes that refer to the same duplication event are identified by MCMCtree as cross-braced (that is, one is chosen as the 'driver' node, the rest are 'mirrored' nodes). In other words, the sampling during the Markov chain Monte Carlo (MCMC) for cross-braced nodes is not independent: the same posterior time density is inferred for matching mirror–driver nodes (see 'Additional methods' for details on our cross-bracing approach).

TIMETREE INFERENCE ANALYSES

Timetree inference with the program MCMCtree (PAML v.4.10.7 (ref. (*93*))) proceeded under both the GBM and ILN relaxed-clock models. We specified a vague rate prior with the shape parameter equal to 2 and the scale parameter equal to 2.5: Γ(2, 2.5). This gamma distribution is meant to account for the uncertainty on our estimate for the mean evolutionary rate, ~0.81 substitutions per site per time unit, which we calculated by dividing the tree height of our best-scoring ML tree (Supplementary Information) into the estimated mean root age of our phylogeny (that is, 4.520 Ga, time unit = 10^9 years; see 'Fossil calibrations' in Supplementary Information for justifications on used calibrations). Given that we are estimating very deep divergences, the molecular clock may be seriously violated. Therefore, we applied a very diffuse gamma prior on the rate variation parameter (σ^2), Γ(1, 10), so that it is centred around $\sigma^2 = 0.1$. To incorporate our uncertainty regarding the tree shape, we specified a uniform kernel density for the birth–death sampling process by setting the birth and death processes to 1, λ (per-lineage birth rate) = μ (per-lineage death rate) = 1, and the sampling frequency to ρ (sampling fraction) = 0.1. Our main analysis consisted of inferring the timetree for the partitioned dataset under both the GBM and the ILN relaxed-clock models in which nodes that correspond to the same divergences are cross-braced (that is, hereby referred to as cross-bracing A). In addition, we ran 10 additional inference analyses to benchmark the effect that partitioning, cross-bracing and relaxed-clock models can have on species divergence time estimation: (1) GBM + concatenated alignment + cross-bracing A, (2) GBM + concatenated alignment + cross-bracing B (only nodes that correspond to the same divergences for which there are fossil constraints are cross-braced), (3) GBM + concatenated alignment + without cross-bracing, (4) GBM + partitioned alignment + cross-bracing B, (5) GBM + partitioned alignment + without cross-bracing, (6) ILN + concatenated alignment + cross-bracing A, (7) ILN + concatenated alignment + cross-bracing B, (8) ILN + concatenated alignment + without cross-bracing, (9) ILN + partitioned alignment + cross-bracing B, and (10) ILN + partitioned alignment + without cross-bracing. Lastly, we used (1) individual gene alignments, (2) a leave-one-out strategy (rate prior changed for alignments without *ATP* and *Leu*, Γ(2, 2.2), and without *Tyr*, Γ(2, 2.3), but was Γ(2, 2.5) for the rest; see 'Additional methods'), and (3) a more complex substitution model (*94*) to assess their impact on timetree inference. Refer to 'Additional methods' for details on how we parsed the dataset we used for timetree inference analyses, ran PAML programs CODEML and MCMCtree to approximate the likelihood calculation (*95*), and carried out the MCMC diagnostics for the results obtained under each of the previously mentioned scenarios.

GENOME SIZE AND CELLULAR FEATURES

We simulated 100 samples of 'KEGG genomes' based on the probabilities of each of the (7,467) gene families being present in LUCA using the random.rand function in numpy (*96*). The mean number of KEGG gene families was 1,298.25, the 95% HPD (highest posterior density) minimum was 1,255 and the maximum was 1,340. To infer the relationship between the number of KEGG KO gene families encoded by a genome, the number of proteins and the genome size, we used LOESS (locally estimated scatter-plot smoothing) regression to estimate the relationship between the number of KOs and (1) the number of protein-coding genes and (2) the genome size for the 700 prokaryotic genomes used in the LUCA reconstruction. To ensure that our inference of genome size is robust to uncertainty in the number of paralogues that can be expected to have been present in LUCA, we used the presence of probability for each of these KEGG KO gene families rather than the estimated copy number. We used the predict function to estimate the protein-coding genes and genome size of LUCA using these models and the simulated gene contents encoded with 95% confidence intervals.

ADDITIONAL METHODS

Cross-bracing approach implemented in MCMCtree

The PAML program MCMCtree was implemented to allow for the analysis of duplicated genes or proteins so that some nodes in the tree corresponding to the same speciation events in different paralogues share the same age. We used the tree topology depicted in Supplementary Fig. 5 to explain how users can label driver or mirror nodes (more on these terms below) so that the program identifies them as sharing the same speciation events. The tree topology shown in Supplementary Fig. 5 can be written in Newick format as:

```
(((A1,A2),A3),((B1,B2),B3));
```

In this example, A and B are paralogues and the corresponding tips labelled as A1–A3 and B1–B3 represent different species. Node *r* represents a duplication event, whereas other nodes are speciation events. If we want to constrain the same speciation events to have the same age (that is, Supplementary Fig. 5, see labels *a* and *b* (that is, A1–A2 ancestor and B1–B2 ancestor, respectively) and labels *v* and *b* (that is, A1–A2–A3 ancestor and B1–B2–B3 ancestor, respectively), we use node labels in the format #1, #2, and so on to identify such nodes:

```
(((A1, A2) #1, A3) #2, ((B1, B2) [#1 B{0.2, 0.4}], B3) #2) 'B(0.9,1.1)';
```

Node *a* and node *b* are assigned the same label (#1) and so they share the same age (t): $t_a = t_b$. Similarly, node *u* and node *v* have the same age: $t_u = t_v$. The former nodes are further constrained by a soft-bound calibration based on the fossil record or geological evidence: $0.2 < t_a = t_b < 0.4$. The latter, however, does not have fossil constraints and thus the only restriction imposed is that both t_u and t_v are equal. Finally, there is another soft-bound calibration on the root age: $0.9 < t_r < 1.1$.

Among the nodes on the tree with the same label (for example, those nodes labelled with #1 and those with #2 in our example), one is chosen as the driver node, whereas the others are mirror nodes. If calibration information is provided on one of the shared nodes (for example, nodes *a* and *b* in Supplementary Fig. 5), the same information therefore applies to all shared nodes. If calibration information is provided on multiple shared nodes, that information has to be the same (for example, you could not constrain node *a* with a different calibration used to constrain node *b* in Supplementary Fig. 5). The time prior (or the prior on all node ages on the tree) is constructed by using a density at the root of the tree, which is specified by the user (for example, 'B(0.9,1.1)' in our example, which has a minimum of 0.9 and a maximum of 1.1). The ages of all non-calibrated nodes are given by the uniform density. This time prior is similar to that used by ref. (*29*). The parameters in the birth–death sampling process (λ, μ, ρ; specified using the option BDparas in the control file that executes MCMCtree) are ignored. It is noteworthy that more than two nodes can have the same label but one node cannot have two or more labels. In addition, the prior on rates does not distinguish between speciation and duplication events. The implemented cross-bracing approach can only be enabled if option duplication = 1 is included in the control file. By default, this option is set to 0 and users are not required to include it in the control file (that is, the default option is duplication = 0).

TIMETREE INFERENCE

Data parsing

Eight paralogues were initially selected based on previous work showing a likely duplication event before LUCA: the amino- and carboxy-terminal regions from carbamoyl phosphate synthetase, aspartate and ornithine transcarbamoylases, histidine biosynthesis genes *A* and *F*, catalytic and non-catalytic subunits from ATP synthase (*ATP*), elongation factor Tu and G (*EF*), signal recognition protein and signal recognition particle receptor (*SRP*), tyrosyl-tRNA and tryptophanyl-tRNA synthetases (*Tyr*), and leucyl- and valyl-tRNA synthetases (*Leu*) (*26*). Gene families were identified using BLASTp (*97*). Sequences were downloaded from NCBI (*98*), aligned with MUSCLE (*99*) and trimmed with TrimAl (*100*) (-strict). Individual gene trees were inferred under the LG+C20+F+G substitution model implemented in IQ-TREE 2 (ref. (*86*)). These trees were manually inspected and curated to remove non-homologous sequences, horizontal gene transfers, exceptionally short or long sequences and extremely long branches. Recent paralogues or taxa of inconsistent and/or uncertain placement inferred with RogueNaRok (*101*) were also removed. Independent verification of an archaeal or bacterial deep split was achieved using minimal ancestor deviation (*102*). This filtering process resulted in the five pairs of paralogous gene families (*27*) (*ATP*, *EF*, *SRP*, *Tyr* and *Leu*) that we used to estimate the origination time of LUCA. The alignment used for timetree inference consisted of 246 species, with the majority of taxa having at least two copies (for some eukaryotes, they may be represented by plastid, mitochondrial and nuclear sequences).

To assess the impact that partitioning can have on divergence time estimates, we ran our inference analyses with both a concatenated and a partitioned alignment (that is, gene

partitioning scheme). We used PAML v.4.10.7 (programs CODEML and MCMCtree) for all divergence time estimation analyses. Given that a fixed tree topology is required for timetree inference with MCMCtree, we inferred the best-scoring ML tree with IQ-TREE 2 under the LG+C20+F+G4 (ref. (*103*)) model following our previous phylogenetic analyses. We then modified the resulting inferred tree topology following consensus views of species-level relationships (*34*, *35*, *104*), which we calibrated with the available fossil calibrations (see below). In addition, we ran three sensitivity tests: timetree inference (1) with each gene alignment separately, (2) under a leave-one-out strategy in which each gene alignment was iteratively removed from the concatenated dataset (for example, remove gene *ATP* but keep genes *EF*, *Leu*, *SRP* and *Tyr* concatenated in a unique alignment block; apply the same procedure for each gene family), and (3) using the vector of branch lengths, the gradient vector and the Hessian matrix estimated under a complex substitution model (bsinBV method described in ref. (*94*)) with the concatenated dataset used for our core analyses. Four of the gene alignments generated for the leave-one-out strategy had gap-only sequences, these were removed when re-inferring the branch lengths under the LG+C20+F+G4 model (that is, without *ATP*, 241 species; without *EF*, 236 species; without *Leu*, 243 species; without *Tyr*, 244 species). We used these trees to set the rate prior used for timetree inference for those alignments not including *ATP*, *EF*, *Leu* or *Tyr*, respectively. The β value (scale parameter) for the rate prior used when analysing alignments without *ATP*, *Leu* and *Tyr* changed minimally but we updated the corresponding rate priors accordingly (see above). When not including *SRP*, the alignment did not have any sequences removed (that is, 246 species). All alignments were analysed with the same rate prior, $\Gamma(2, 2.5)$, except for the three previously mentioned alignments.

A

Approximating the likelihood calculation during timetree inference using PAML programs

Before timetree inference, we ran the CODEML program to infer the branch lengths of the fixed tree topology, the gradient (first derivative of the likelihood function) and the Hessian matrix (second derivative of the likelihood function); the vectors and matrix are required to approximate the likelihood function in the dating program MCMCtree (*95*), an approach that substantially reduces computational time (*105*). Given that CODEML does not implement the CAT (Bayesian mixture model for across-site heterogeneity) model, we ran our analyses under the closest available substitution model: LG+F+G4 (model = 3). We calculated the aforementioned vectors and matrix for each of the five gene alignments (that is, required for the partitioned alignment), for the concatenated alignment and for the concatenated alignments used for the leave-one-out strategy; the resulting values are written out in an output file called rst2. We appended the rst2 files generated for each of the five individual alignments in the same order the alignment blocks appear in the partitioned alignment file (for example, the first alignment block corresponds to the *ATP* gene alignment, and thus the first rst2 block will be the one generated when analysing the *ATP* gene alignment with CODEML). We named this file in_5parts.BV. There is only one rst2 output file for the concatenated alignments, which we renamed in.BV (main concatenated alignment and concatenated alignments under

leave-one-out strategy). When analysing each gene alignment separately, we renamed the rst2 files generated for each gene alignment as in.BV.

MCMC diagnostics

All the chains that we ran with MCMCtree for each type of analysis underwent a protocol of MCMC diagnostics consisting of the following steps: (1) flagging and removal of problematic chains; (2) generating convergence plots before and after chain filtering; (3) using the samples collected by those chains that passed the filters (that is, assumed to have converged to the same target distribution) to summarize the results; (4) assessing chain efficiency and convergence by calculating statistics such as R-hat, tail-ESS and bulk-ESS (in-house wrapper function calling Rstan functions, Rstan v.2.21.7; https://mc-stan.org/rstan/); and (5) generating the timetrees for each type of analysis with confidence intervals and high-posterior densities to show the uncertainty surrounding the estimated divergence times. Tail-ESS is a diagnostic tool that we used to assess the sampling efficiency in the tails of the posterior distributions of all estimated divergence times, which corresponds to the minimum of the effective sample sizes for quantiles 2.5% and 97.5%. To assess the sampling efficiency in the bulk of the posterior distributions of all estimated divergence, we used bulk-ESS, which uses rank-normalized draws. Note that if tail-ESS and bulk-ESS values are larger than 100, the chains are assumed to have been efficient and reliable parameter estimates (that is, divergence times in our case). R-hat is a convergence diagnostic measure that we used to compare between- and within-chain divergence time estimates to assess chain mixing. If R-hat values are larger than 1.05, between- and within-chain estimates do not agree and thus mixing has been poor. Lastly, we assessed the impact that truncation may have on the estimated divergence times by running MCMCtree when sampling from the prior (that is, the same settings specified above but without using sequence data, which set the prior distribution to be the target distribution during the MCMC). To summarize the samples collected during this analysis, we carried out the same MCMC diagnostics procedure previously mentioned. Supplementary Fig. 6 shows our calibration densities (commonly referred to as user-specified priors, see justifications for used calibrations above) versus the marginal densities (also known as effective priors) that MCMCtree infers when building the joint prior (that is, a prior built without sequence data that considers age constraints specified by the user, the birth–death with sampling process to infer the time densities for the uncalibrated nodes, the rate priors, and so on). We provide all our results for these quality-control checks in our GitHub repository (https://github.com/sabifo4/LUCA-divtimes) and in Extended Data Fig. 1, Supplementary Figs. 7–10 and Supplementary Data 6. Data, figures and tables used and/or generated following a step-by-step tutorial are detailed in the GitHub repository for each inference analysis.

Additional sensitivity analyses

We compared the divergence times we estimated with the concatenated dataset under the calibration strategy cross-bracing A with those inferred (1) for each gene, (2) for gene alignments analysed under a leave-one-out strategy, and (3) for the main concatenated dataset but when using the vector of branch lengths, the gradient vector and the Hessian

matrix estimated under a more complex substitution model (*94*). The results are summarized in Extended Data Fig. 2 and Supplementary Data 7 and 8. The same pattern regarding the calibration densities and marginal densities when the tree topology was pruned (that is, see above for details on the leave-one-out strategy) was observed, and thus no additional figures have been generated. As for our main analyses, the results for these additional sensitivity analyses can be found on our GitHub repository (https://github.com/sabifo4/LUCA-divtimes).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

DATA AVAILABILITY

All data required to interpret, verify and extend the research in this article can be found at our figshare repository at https://doi.org/10.6084/m9.figshare.24428659 (ref. (*106*)) for the reconciliation and phylogenomic analyses and GitHub at https://github.com/sabifo4/LUCA-divtimes (ref. (*107*)) for the molecular clock analyses. Additional data are available at the University of Bristol data repository, data.bris, at https://doi.org/10.5523/bris.405xnm7ei36d2cj65nrirg3ip (ref. (*108*)).

CODE AVAILABILITY

All code relating to the dating analysis can be found on GitHub at https://github.com/sabifo4/LUCA-divtimes (ref. (*107*)), and other custom scripts can be found in our figshare repository at https://doi.org/10.6084/m9.figshare.24428659 (ref. (*106*)).

ACKNOWLEDGEMENTS

Our research is funded by the John Templeton Foundation (62220 to P.C.J.D., N.L., T.M.L., D.P., G.A.S., T.A.W. and Z.Y.; the opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation), Biotechnology and Biological Sciences Research Council (BB/T012773/1 to P.C.J.D. and Z.Y.; BB/T012951/1 to Z.Y.), by the European Research Council under the European Union's Horizon 2020 research and innovation programme (947317 ASymbEL to A.S.; 714774, GENECLOCKS to G.J.S.), Leverhulme Trust (RF-2022-167 to P.C.J.D.), Gordon and Betty Moore Foundation (GBMF9741 to T.A.W., D.P., P.C.J.D., A.S. and G.J.S.; GBMF9346 to A.S.), Royal Society (University Research Fellowship (URF) to T.A.W.), the Simons Foundation (735929LPI to A.S.) and the University of Bristol (University Research Fellowship (URF) to D.P.).

AUTHOR CONTRIBUTIONS

The project was conceived and designed by P.C.J.D., T.M.L., D.P., G.J.S., A.S. and T.A.W. Dating analyses were performed by H.C.B., J.W.C., S.Á.C., P.J.C.D. and E.R.R.M. T.A.M., N.D. and E.R.R.M. performed single-copy orthologue analysis for species-tree inference. L.L.S., G.J.S., T.A.W. and E.R.R.M. performed reconciliation analysis. E.R.R.M. performed homologous gene family annotation, sequence, alignment, gene tree inference and sensitivity tests. E.R.R.M., A.S. and T.A.W. performed metabolic analysis and interpretation. T.M.L., S.D. and R.A.B. provided biogeochemical interpretation. E.R.R.M., T.M.L., A.S., T.A.W., D.P. and P.J.C.D. drafted the article to which all authors (including X.C., N.L., Z.Y. and G.A.S.) contributed.

ADDITIONAL INFORMATION

Extended data is available for this paper at https://doi.org/10.1038/s41559-024-02461-1.
All Extended data figures can be accessed here:
https://www.nature.com/articles/s41559-024-02461-1#Sec22

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-024-02461-1.

All Supplementary materials can be accessed here*:*
https://www.nature.com/articles/s41559-024-02461-1#Sec23

Correspondence and requests for materials should be addressed to Edmund R. R. Moody, Davide Pisani, Tom A. Williams, Timothy M. Lenton or Philip C. J. Donoghue.

REFERENCES

1. D. L. Theobald, A formal test of the theory of universal common ancestry. *Nature* **465**, 219–222 (2010).

2. C. R. Woese, G. E. Fox, The concept of cellular evolution. *J. Mol. Evol.* **10**, 1–6 (1977).

3. B. G. Mirkin, T. I. Fenner, M. Y. Galperin, E. V. Koonin, Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* **3**, 2 (2003).

4. C. A. Ouzounis, V. Kunin, N. Darzentas, L. Goldovsky, A minimal estimate for the gene content of the last universal common ancestor--exobiology from a terrestrial perspective. *Res. Microbiol.* **157**, 57–68 (2006).

5. J. P. Gogarten, D. Deamer, Is LUCA a thermophilic progenote?, *Nature microbiology*. **1** (2016)p. 16229.

6. M. C. Weiss, F. L. Sousa, N. Mrnjavac, S. Neukirchen, M. Roettger, S. Nelson-Sathi, W. F. Martin, The physiology and habitat of the last universal common ancestor. *Nat Microbiol* **1**, 16116 (2016).

7. A. J. Crapitto, A. Campbell, A. J. Harris, A. D. Goldman, A consensus view of the proteome of the last universal common ancestor. *Ecol. Evol.* **12**, e8930 (2022).

8. N. Kyrpides, R. Overbeek, C. Ouzounis, Universal protein families and the functional content of the last universal common ancestor. *J. Mol. Evol.* **49**, 413–423 (1999).

9. E. V. Koonin, Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.* **1**, 127–136 (2003).

10. J. K. Harris, S. T. Kelley, G. B. Spiegelman, N. R. Pace, The genetic core of the universal ancestor. *Genome Res.* **13**, 407–412 (2003).

11. E. J. Javaux, Challenges in evidencing the earliest traces of life. *Nature* **572**, 451–460 (2019).

12. K. Lepot, Signatures of early microbial life from the Archean (4 to 2.5 Ga) eon. *Earth Sci. Rev.* **209**, 103296 (2020).

13. H. C. Betts, M. N. Puttick, J. W. Clark, T. A. Williams, P. C. J. Donoghue, D. Pisani, Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nat Ecol Evol* **2**, 1556–1562 (2018).

14. Q. Zhu, U. Mai, W. Pfeiffer, S. Janssen, F. Asnicar, J. G. Sanders, P. Belda-Ferre, G. A. Al-Ghalith, E. Kopylova, D. McDonald, T. Kosciolek, J. B. Yin, S. Huang, N. Salam, J.-Y. Jiao, Z. Wu, Z. Z. Xu, K. Cantrell, Y. Yang, E. Sayyari, M. Rabiee, J. T. Morton, S. Podell, D. Knights, W.-J. Li, C. Huttenhower, N. Segata, L. Smarr, S. Mirarab, R. Knight, Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477 (2019).

15. E. R. R. Moody, T. A. Mahendrarajah, N. Dombrowski, J. W. Clark, C. Petitjean, P. Offre, G. J. Szöllősi, A. Spang, T. A. Williams, An estimate of the deepest branches of the tree of life from ancient vertically evolving genes. *Elife* **11** (2022).

16. R. M. Schwartz, M. O. Dayhoff, Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. *Science* **199**, 395–403 (1978).

17. P. M. Shih, N. J. Matzke, Primary endosymbiosis events date to the later Proterozoic with cross-calibrated phylogenetic dating of duplicated ATPase proteins. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 12355–12360 (2013).

18. T. A. Mahendrarajah, E. R. R. Moody, D. Schrempf, L. L. Szánthó, N. Dombrowski, A. A. Davín, D. Pisani, P. C. J. Donoghue, G. J. Szöllősi, T. A. Williams, A. Spang, ATP synthase evolution on a cross-braced dated tree of life. *Nat. Commun.* **14**, 7456 (2023).

19. W. F. Bottke, M. D. Norman, The Late Heavy Bombardment. *Annu. Rev. Earth Planet. Sci.* **45**, 619–647 (2017).

20. J. Reimink, C. Crow, D. Moser, B. Jacobsen, A. Bauer, T. Chacko, Quantifying the effect of late bombardment on terrestrial zircons. *Earth Planet. Sci. Lett.* **604**, 118007 (2023).

21. P. Boehnke, T. M. Harrison, Illusory late heavy bombardments. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 10802–10806 (2016).

22. G. Ryder, Mass flux in the ancient Earth-Moon system and benign implications for the origin of life on Earth. *J. Geophys. Res.* **107** (2002).

23. W. K. Hartmann, History of the terminal cataclysm paradigm: Epistemology of a planetary bombardment that never (?) happened. *Geosciences (Basel)* **9**, 285 (2019).

24. N. J. Planavsky, D. Asael, A. Hofmann, C. T. Reinhard, S. V. Lalonde, A. Knudsen, X. Wang, F. Ossa Ossa, E. Pecoits, A. J. B. Smith, N. J. Beukes, A. Bekker, T. M. Johnson, K. O. Konhauser, T. W. Lyons, O. J. Rouxel, Evidence for oxygenic photosynthesis half a billion years before the Great Oxidation Event. *Nat. Geosci.* **7**, 283–286 (2014).

25. F. Ossa Ossa, A. Hofmann, J. E. Spangenberg, S. W. Poulton, E. E. Stüeken, R. Schoenberg, B. Eickmann, M. Wille, M. Butler, A. Bekker, Limited oxygen production in the Mesoarchean ocean. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 6647–6652 (2019).

26. S. B. Mukasa, A. H. Wilson, K. R. Young, Geochronological constraints on the magmatic and tectonic development of the Pongola Supergroup (Central Region), South Africa. *Precambrian Res.* **224**, 268–286 (2013).

27. O. Zhaxybayeva, P. Lapierre, J. P. Gogarten, Ancient gene duplications and the root(s) of the tree of life. *Protoplasma* **227**, 53–64 (2005).

28. P. C. J. Donoghue, Z. Yang, The evolution of methods for establishing evolutionary timescales. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **371**, 20160020 (2016).

29. J. L. Thorne, H. Kishino, I. S. Painter, Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* **15**, 1647–1657 (1998).

30. Z. Yang, B. Rannala, Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol. Biol. Evol.* **23**, 212–226 (2006).

31. B. Rannala, Z. Yang, Inferring speciation times under an episodic molecular clock. *Syst. Biol.* **56**, 453–466 (2007).

32. P. Lemey, A. Rambaut, J. J. Welch, M. A. Suchard, Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27**, 1877–1885 (2010).

33. J. M. Craig, S. Kumar, S. B. Hedges, The origin of eukaryotes and rise in complexity were synchronous with the rise in oxygen. *Front. Bioinform.* **3**, 1233281 (2023).

34. M. Aouad, J.-P. Flandrois, F. Jauffrit, M. Gouy, S. Gribaldo, C. Brochier-Armanet, A divide-and-conquer phylogenomic approach based on character supermatrices resolves early steps in the evolution of the Archaea. *BMC Ecol. Evol.* **22**, 1 (2022).

35. G. A. Coleman, A. A. Davín, T. A. Mahendrarajah, L. L. Szánthó, A. Spang, P. Hugenholtz, G. J. Szöllősi, T. A. Williams, A rooted phylogeny resolves early bacterial evolution. *Science* **372** (2021).

36. L. Guy, T. J. G. Ettema, The archaeal "TACK" superphylum and the origin of eukaryotes. *Trends Microbiol.* **19**, 580–587 (2011).

37. A. Spang, J. H. Saw, S. L. Jørgensen, K. Zaremba-Niedzwiedzka, J. Martijn, A. E. Lind, R. van Eijk, C. Schleper, L. Guy, T. J. G. Ettema, Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).

38. K. Zaremba-Niedzwiedzka, E. F. Caceres, J. H. Saw, D. Bäckström, L. Juzokaite, E. Vancaester, K. W. Seitz, K. Anantharaman, P. Starnawski, K. U. Kjeldsen, M. B. Stott, T. Nunoura, J. F. Banfield, A. Schramm, B. J. Baker, A. Spang, T. J. G. Ettema, Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).

39. L. Eme, D. Tamarit, E. F. Caceres, C. W. Stairs, V. De Anda, M. E. Schön, K. W. Seitz, N. Dombrowski, W. H. Lewis, F. Homa, J. H. Saw, J. Lombard, T. Nunoura, W.-J. Li, Z.-S. Hua, L.-X. Chen, J. F. Banfield, E. S. John, A.-L. Reysenbach, M. B. Stott, A. Schramm, K. U. Kjeldsen, A. P. Teske, B. J. Baker, T. J. G. Ettema, Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature* **618**, 992–999 (2023).

40. K. Raymann, C. Brochier-Armanet, S. Gribaldo, The two-domain tree of life is linked to a new root for the Archaea. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 6670–6675 (2015).

41. D. Megrian, N. Taib, A. L. Jaffe, J. F. Banfield, S. Gribaldo, Ancient origin and constrained evolution of the division and cell wall gene cluster in Bacteria. *Nat. Microbiol.* **7**, 2114–2127 (2022).

42. C. T. Brown, L. A. Hug, B. C. Thomas, I. Sharon, C. J. Castelle, A. Singh, M. J. Wilkins, K. C. Wrighton, K. H. Williams, J. F. Banfield, Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).

43. C. Rinke, P. Schwientek, A. Sczyrba, N. N. Ivanova, I. J. Anderson, J.-F. Cheng, A. Darling, S. Malfatti, B. K. Swan, E. A. Gies, J. A. Dodsworth, B. P. Hedlund, G. Tsiamis, S. M. Sievert, W.-T. Liu, J. A. Eisen, S. J. Hallam, N. C. Kyrpides, R. Stepanauskas, E. M. Rubin, P. Hugenholtz, T. Woyke, Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).

44. H. Shimodaira, An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508 (2002).

45. N. Taib, D. Megrian, J. Witwinowski, P. Adam, D. Poppleton, G. Borrel, C. Beloin, S. Gribaldo, Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. *Nat Ecol Evol* **4**, 1661–1672 (2020).

46. G. J. Szöllősi, W. Rosikiewicz, B. Boussau, E. Tannier, V. Daubin, Efficient exploration of the space of reconciled gene trees. *Syst. Biol.* **62**, 901–912 (2013).

47. M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

48. T. A. Williams, G. J. Szöllősi, A. Spang, P. G. Foster, S. E. Heaps, B. Boussau, T. J. G. Ettema, T. M. Embley, Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E4602–E4611 (2017).

49. J. E. Dharamshi, S. Köstlbacher, M. E. Schön, A. Collingro, T. J. G. Ettema, M. Horn, Gene gain facilitated endosymbiotic evolution of Chlamydiae. *Nat. Microbiol.* **8**, 40–54 (2023).

50. W. F. Doolittle, Phylogenetic classification and the universal tree. *Science* **284**, 2124–2129 (1999).

51. T. Dagan, W. Martin, The tree of one percent. *Genome Biol.* **7**, 118 (2006).

52. R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, D. A. Natale, The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).

53. S. W. Ragsdale, E. Pierce, Acetogenesis and the Wood–Ljungdahl pathway of CO2 fixation. *Biochim. Biophys. Acta Proteins Proteom.* **1784**, 1873–1898 (2008).

54. K. Schuchmann, V. Müller, Autotrophy at the thermodynamic limit of life: a model for energy conservation in acetogenic bacteria. *Nat. Rev. Microbiol.* **12**, 809–821 (2014).

55. K. Schuchmann, V. Müller, Energetics and application of heterotrophy in acetogenic bacteria. *Appl. Environ. Microbiol.* **82**, 4056–4069 (2016).

56. N. Iwabe, K. Kuma, M. Hasegawa, S. Osawa, T. Miyata, Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 9355–9359 (1989).

57. J. P. Gogarten, H. Kibak, P. Dittrich, L. Taiz, E. J. Bowman, B. J. Bowman, M. F. Manolson, R. J. Poole, T. Date, T. Oshima, J. Konishi, K. Denda, M. Yoshida, Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 6661–6665 (1989).

58. E. V. Koonin, K. S. Makarova, Origins and evolution of CRISPR-Cas systems. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20180087 (2019).

59. M. Krupovic, V. V. Dolja, E. V. Koonin, The LUCA and its complex virome. *Nat. Rev. Microbiol.* **18**, 661–670 (2020).

60. E. V. Koonin, V. V. Dolja, M. Krupovic, The logic of virus evolution. *Cell Host Microbe* **30**, 917–929 (2022).

61. M. A. Lever, Acetogenesis in the energy-starved deep biosphere – A paradox? *Front. Microbiol.* **2** (2012).

62. W. Martin, M. J. Russell, On the origin of biochemistry at an alkaline hydrothermal vent. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **362**, 1887–1925 (2007).

63. R. J. Catchpole, P. Forterre, The evolution of reverse gyrase suggests a nonhyperthermophilic last universal common ancestor. *Mol. Biol. Evol.* **36**, 2737–2747 (2019).

64. M. Groussin, B. Boussau, S. Charles, S. Blanquart, M. Gouy, The molecular signal for the adaptation to cold temperature during early life on Earth. *Biol. Lett.* **9**, 20130608 (2013).

65. B. Boussau, S. Blanquart, A. Necsulea, N. Lartillot, M. Gouy, Parallel adaptations to high temperatures in the Archaean eon. *Nature* **456**, 942–945 (2008).

66. A. Chandor, O. Berteau, T. Douki, D. Gasparutto, Y. Sanakis, S. Ollagnier-de-Choudens, M. Atta, M. Fontecave, Dinucleotide spore photoproduct, a minimal substrate of the DNA repair spore photoproduct lyase enzyme from Bacillus subtilis. *J. Biol. Chem.* **281**, 26922–26931 (2006).

67. T. Chandra, S. C. Silver, E. Zilinskas, E. M. Shepard, W. E. Broderick, J. B. Broderick, Spore photoproduct lyase catalyzes specific repair of the 5R but not the 5S spore photoproduct. *J. Am. Chem. Soc.* **131**, 2420–2421 (2009).

68. J. F. Kasting, The evolution of the prebiotic atmosphere. *Origins Life Evol Biosphere* **14**, 75–82 (1984).

69. P. A. Kharecha, "A Coupled Atmosphere–Ecosystem Model of the Early Archean Biosphere," thesis, Pennsylvania State (2005).

70. P. Barth, E. E. Stüeken, C. Helling, L. Rossmanith, Y. Peng, W. Walters, M. Claire, Isotopic constraints on lightning as a source of fixed nitrogen in Earth's early biosphere. *Nat. Geosci.* **16**, 478–484 (2023).

71. F. Tian, J. F. Kasting, K. Zahnle, Revisiting HCN formation in Earth's early atmosphere. *Earth Planet. Sci. Lett.* **308**, 417–423 (2011).

72. K. J. Zahnle, Photochemistry of methane and the formation of hydrocyanic acid (HCN) in the Earth's early atmosphere. *J. Geophys. Res.* **91**, 2819–2834 (1986).

73. E. E. Stüeken, T. Boocock, K. Szilas, S. Mikhail, N. J. Gardiner, Reconstructing nitrogen sources to Earth's earliest biosphere at 3.7 Ga. *Front. Earth Sci.* **9** (2021).

74. F. D. Ciccarelli, T. Doerks, C. von Mering, C. J. Creevey, B. Snel, P. Bork, Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287 (2006).

75. N. Yutin, K. S. Makarova, S. L. Mekhedov, Y. I. Wolf, E. V. Koonin, The deep archaeal roots of eukaryotes. *Mol. Biol. Evol.* **25**, 1619–1630 (2008).

76. C. Petitjean, P. Deschamps, P. López-García, D. Moreira, Rooting the domain archaea by phylogenomic analysis supports the foundation of the new kingdom Proteoarchaeota. *Genome Biol. Evol.* **7**, 191–204 (2014).

77. T. A. Williams, C. J. Cox, P. G. Foster, G. J. Szöllősi, T. M. Embley, Phylogenomics provides robust support for a two-domains tree of life. *Nat Ecol Evol* **4**, 138–147 (2020).

78. C. Rinke, M. Chuvochina, A. J. Mussig, P.-A. Chaumeil, A. A. Davín, D. W. Waite, W. B. Whitman, D. H. Parks, P. Hugenholtz, A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nat Microbiol* **6**, 946–959 (2021).

79. D. H. Parks, M. Chuvochina, P.-A. Chaumeil, C. Rinke, A. J. Mussig, P. Hugenholtz, Selection of representative genomes for 24,706 bacterial and archaeal species clusters provide a complete genome-based taxonomy, *bioRxiv* (2019). https://doi.org/10.1101/771964.

80. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29-37 (2011).

81. K. S. Makarova, Y. I. Wolf, E. V. Koonin, Archaeal clusters of Orthologous Genes (arCOGs): An update and application for analysis of shared features between Thermococcales, methanococcales, and Methanobacteriales. *Life (Basel)* **5**, 818–840 (2015).

82. D. H. Haft, J. D. Selengut, O. White, The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373 (2003).

83. K. Katoh, K.-I. Kuma, H. Toh, T. Miyata, MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).

84. K. Katoh, K. Misawa, K.-I. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).

85. A. Criscuolo, S. Gribaldo, BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).

86. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, R. Lanfear, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

87. S. R. Eddy, Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).

88. E. Susko, A. J. Roger, On reduced amino acid alphabets for phylogenetic inference. *Mol. Biol. Evol.* **24**, 2139–2150 (2007).

89. C. P. Cantalapiedra, A. Hernández-Plaza, I. Letunic, P. Bork, J. Huerta-Cepas, Eggnog-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).

90. B. Buchfink, K. Reuter, H.-G. Drost, Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).

91. G. J. Szöllősi, A. A. Davín, E. Tannier, V. Daubin, B. Boussau, Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140335 (2015).

92. A. M. Eren, E. Kiefl, A. Shaiber, I. Veseli, S. E. Miller, M. S. Schechter, I. Fink, J. N. Pan, M. Yousef, E. C. Fogarty, F. Trigodet, A. R. Watson, Ö. C. Esen, R. M. Moore, Q. Clayssen, M. D. Lee, V. Kivenson, E. D. Graham, B. D. Merrill, A. Karkman, D. Blankenberg, J. M. Eppley, A. Sjödin, J. J. Scott, X. Vázquez-Campos, L. J. McKay, E. A. McDaniel, S. L. R. Stevens, R. E. Anderson, J. Fuessel, A. Fernandez-Guerra, L. Maignien, T. O. Delmont, A. D. Willis, Community-led, integrated, reproducible multi-omics with anvi'o. *Nat. Microbiol.* **6**, 3–6 (2021).

93. Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

94. S. Wang, H. Luo, Dating the bacterial tree of life based on ancient symbiosis, *bioRxiv* (2023). https://doi.org/10.1101/2023.06.18.545440.

95. M. dos Reis, Z. Yang, Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. *Mol. Biol. Evol.* **28**, 2161–2172 (2011).

96. C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. Del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T. E. Oliphant, Array programming with NumPy. *Nature* **585**, 357–362 (2020).

97. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

98. E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, J. Ye, Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **39**, D38-51 (2011).

99. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

100. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

101. A. Aberer, D. Krompaß, A. Stamatakis, RogueNaRok: An efficient and exact algorithm for rogue taxon identification. *Heidelberg Institute for Theoretical Studies: Exelixis-RRDR-2011--10* (2011).

102. F. D. K. Tria, G. Landan, T. Dagan, Phylogenetic rooting using minimal ancestor deviation. *Nat Ecol Evol* **1**, 193 (2017).

103. D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

104. F. Burki, A. J. Roger, M. W. Brown, A. G. B. Simpson, The new tree of eukaryotes. *Trends Ecol. Evol.* **35**, 43–55 (2020).

105. F. U. Battistuzzi, P. Billing-Ross, A. Paliwal, S. Kumar, Fast and slow implementations of relaxed-clock methods show similar patterns of accuracy in estimating divergence times. *Mol. Biol. Evol.* **28**, 2439–2442 (2011).

106. E. Moody, The nature of the Last Universal Common Ancestor and its impact on the early Earth system, figshare (2024); https://doi.org/10.6084/M9.FIGSHARE.24428659.

107. A. C. Sandra, *Sabifo4/LUCA-Divtimes: V1.0.0* (Zenodo, 2024; http://dx.doi.org/10.5281/ZENODO.11260523).

108. E. Moody, S. A. Carretero, T. A. Mahendrarajah, J. W. Clark, H. C. Betts, N. Dombrowski, L. Szantho, R. Boyle, S. Daines, X. Chen, N. Lane, Z. Yang, G. A. Shields, G. J. Szöllősi, A. Spang, D. Pisani, T. Williams, T. Lenton, P. Donoghue, Data from Moody et al. (2024) The nature of the Last Universal Common Ancestor and its impact on the early Earth system. Nature Ecology and Evolution, University of Bristol (2024); https://doi.org/10.5523/BRIS.405XNM7EI36D2CJ65NRIRG3IP.

109. Y. Darzi, I. Letunic, P. Bork, T. Yamada, iPath3.0: interactive pathways explorer v3. *Nucleic Acids Res.* **46**, W510–W513 (2018).

APPENDIX B

Acknowledgements
Curriculum vitae
List of publications

ACKNOWLEDGEMENTS

In many ways, this thesis isn't just about me. It's the culmination of every person who rode alongside me on this journey, through the highs and the lows and everything else in between. I truly believe this would not have been a reality without all of you and I want to take a moment to extend my deepest appreciation to you all.

I am what academia might call a "non-traditional" student… I didn't follow a straight path from my undergraduate work to this PhD and, instead, spent time in between working and chasing other endeavors. I firmly believe that I couldn't have made it here without the encouragement and support of past supervisors and lab mates who challenged that traditional notion and pushed me to continue on this path despite my age or experience. This began *way* back at UMass (Class of 2012). **Jeff**, thank you for always believing I'd make a good scientist, even when I didn't fully understand what that meant. And of course, to **Amy**, the then PhD-student who patiently trained me and cheered me on in everything I did. I have always admired your tenacity, determination, and courage to return to school after many years to obtain a PhD – it was you who showed me this could be a reality. At the same time, I want to thank those of you in Los Angeles and at CSU Northridge who encouraged me onto this path – **Rachel, Dr. Flores** (and **Sara**), **Alex**, **Cristi**, **Archana**, **David**, **Anthony**, **Michael**, **Chris**, **Jamie**, and **The Perera Family**.

Anja, you are central to this story. I want to extend my sincerest gratitude to you for offering me this position and for your unwavering support in basically everything I've put my mind to. You have challenged me to do my best work, helped me when I struggled, and guided me when I was lost. Beyond that your unflinching faith in my abilities gave me the confidence I needed when self-doubt crept in and helped keep me grounded. You value my opinions and perspectives and provided a safe environment to collaborate in science and with other interests. It's truly amazing to reflect on the past five years and see how much we've both grown, together but in our own separate ways. You have been with me through some of the highest highs of my life and some of the lowest lows, and you have always put me first as a person, having empathy and genuine concern and knowing when to advocate for me when necessary. You have been understanding, patient, and kind, which made the transition into an unfamiliar field much less daunting. Your encouragement has given me the confidence to explore my research, personal, and professional interests. It has been an immense honor to be your first PhD student, and I am proud of how far we have both come.

Next, I would like to also extend my appreciation to **Stefan** and **Laura**, you both played important roles in the progression of my PhD work and supporting me, personally and professionally, as I became a parent. I am grateful that both of you were generous and accommodating, I truly believe that relieved a lot of stress and allowed me to be more present for the earliest moments of Bodhi's life. I will never forget what that grace and kindness afforded me. Along those lines, I want to thank everyone at **HRM**, especially **Marjolijn** and **Mylene**, for handling the administrative aspects of my maternity leave, residence permits, and

contract extensions. And a big thank you to the **IT department**, especially **Mike** and **Sander**, I literally wouldn't have had a machine to write my thesis on without you two. And a special thanks to **Daan** who made time to help with the Dutch translation of my thesis summary.

A heartfelt thank you to members of the **MMB Department** who listened to my presentations, read my papers, provided valuable feedback and suggestions, or were just there to chat. I often felt on the outskirts of our central research themes of the department but felt that my research and my efforts were valued by my colleagues. And although I barely used the lab facilities, I appreciate all the **lab support staf**f. A very special thank you to **Ilsa**, you were one of the first people I got to know once moving to the Netherlands. It had only been two months since arriving here that I hopped on a train with you go to Regensburg for a cultivation course. It is still one of the most memorable trips of my life, and I fondly look back on our time walking to and from the lab, getting pretzels in the morning, and sitting on the playground eating my very first kinderegg.

To **Su** and **Chunjing**. I will relish all the moments we spent together sharing food and stories, our happiness and sadness, playing Tricky towers, and cuddling our cats, although I'm not convinced Naofu was that excited to hang out with me that often. I am so proud of the people that you have both become in the years that I have known you and I happily look forward to sharing a meal with you and your families in China. I could not have done any of this without your support, friendship, and love. To **Annika** and **Sofia**, thank you both for sharing your home and your happiness with me, it has been a joy to watch both of you grow and it was fun to have someone to play Animal Crossing with. **Diana**, **Alejandro**, and **Iara**, thank you for being great friends (and colleagues) and providing so much advice about navigating childbirth in the Netherlands. **Bastiaan**, you are one of the kindest and most enthusiastic people I've ever met, I am lucky to be able to work with you. **Merve**, this experience would have felt a lot lonelier without your courage, honesty, and commitment to your principles – I am proud to know you.

Jamie, this is the second time you are listed in here because our journey goes back to our time at CSU Northridge. As another non-traditional student making a career change, your commitment to your work and social justice helped give me the confidence (then and now) to take risks and have the courage to make changes in my life or the lives of people around me. **Amin**, I truly believe a friendship as smooth and easy as ours is one that will stand the test of time. I am proud of all that you have accomplished since coming here and I am looking forward to seeing you succeed in your new role. To **George**. I can still remember that first lunch I joined you because you were sitting alone in the canteen. Who would've known we'd be here, five years later living in different places but still finding ways back to each other. I am proud of you for finding a career and pathway that suits what you want and need in your life. **Stanley** (and Mochi), I am so very grateful we connected, you have no idea how much joy and laughter you bring to our lives, it has been an honor to share part of this journey with you.

I am grateful for the shared experiences, joy, and care from many other colleagues including: **Laura** (Pacho Sampedro), **Szabina**, **Anandi**, **Emna**, **Annalisa**, **Lia**, **Tom**, **Saara**, **Kirsten**, **Faye**, **Peter**, **Charlotte**, **Rachel**, **Monique**, **Jort**, **Sharon**, **Philip**, **Jessica**, and **Edwin**. I would like to express my immense gratitude for all the **cleaners** and **lunch staff** who have kept our work areas neat and served us food and maintained the coffee machines.

To my paranymphs, **Nina** and **Josh**, I would have been totally lost without you two. I feel like there's a mountain of things I could say about our friendship, but one of the things I appreciate the most is how consistent it is. Even with changing jobs and major life transitions, you're both still there as constants reassuring me and keeping me grounded. I'm proud of us for staying the course for four years, which have felt especially difficult at moments – thank you for always listening to the light stuff and the heavy stuff too and being there to comfort me when I needed it. You have both been voices of reason, guidance, and healthy levels of judgement and criticism mixed in with some serious cheerleading. I have learned so much from both of you and am happy to have had the opportunity to work alongside you. You've been a sounding board for my ideas, writing, frustrations, and just random fleeting thoughts, and I'm confident it has made me a better writer and scientist. But most importantly, like Anja, you both championed (and pushed me) to live my most authentic life outside of my research – you supported me through so many life-changing events, like buying a house, getting cats, having a child, and never once planted a seed of doubt in my mind that I could do this all, and I am forever grateful for that. Nina, we've known each other since the very beginning… we've shared an office, projects, methods, cat food, and so very many cups of coffee. I can't even fathom how many walks we've gone on together in the past 5 years. You made the transition into a new field much less daunting for me and you helped me believe in myself in ways that I really couldn't have imagined. I do genuinely miss working with you, but I am grateful you are doing a job that brings you joy. Josh, thanks for always jumping on board with my social activities, and for talking about sports when no one else wants to. Your commitment to your work inspires me and I honestly hope I can be half the scientist you are.

Wen-Cong, you and I were the only two PhD students in what seemed like a sea of post-docs in our research group. I am happy that I was able to share that experience with you. I am proud of the scientist you are becoming, and you challenge me every day. It is an honor to collaborate with you. To the rest of the Spang-group, past and present: **Carlos**, **Florian**, **Oleks**, **Dina**, **Kim**, **Josje**, **Gerben**, and **Jun-Hoe**. I feel fortunate to work with and learn from all of you. Individually you each bring so many diverse experiences and expertise to our group that helps build a solid, open, collaborative, and welcoming atmosphere. It has truly been great times with you all.

All the work in this thesis would have been impossible without all collaborators who made these studies a reality. A special thanks to: **Tom**, **Ed**, **Gareth**, **Gergely**, **Lénárd**, **Phil** (Hugenholtz), **Adri**, **Davide**, **Phil** (Donoghue), **Courtney**, and **Pierre**.

I'd also like to extend my heartfelt gratitude to the members of my **dissertation assessment committee**, for taking the time to read, evaluate, and assess this thesis and for participating in this defense ceremony.

Life outside of NIOZ has also been rich with friendship, and I have been fortunate enough to build a small but loving and vibrant community here in Noord Holland. To **De Van Hogendorpjes**, it is such a pleasure to be your neighbor, the street is welcoming and it's nice to see your smiling faces. **Carin**, I am so grateful for you, not just for making space to babysit Bodhi, but also for being a kind and charismatic neighbor who brings people together. You are a force of nature, and I admire your strength and courage – I hope I can live with as much grace as you do. I also want to thank **Nathanael**, **Benjamin**, **Phileine**, **Juda**, and **Davi Jack** for sharing your toys, playing with Bodhi, and filling our street with laughter and joy. The utmost gratitude to **Debra** (and **Niel**). Debra, thank you for being so wonderful with Bodhi and having him look forward to spending time with you during the day. You are an excellent teacher, and he has learned so much from you. You were instrumental in giving me some extra little moments to decompress while I was writing my thesis and cheered me on the entire way. I am grateful for your level-headedness, empathy, and willingness to listen when I need to air out my frustrations. And thank you for indulging in my weird interests and hobbies, your courage to take risks and try new things inspires me to do the same. **Joke** and **Bernard**, thank you for being so kind and helpful on this journey and for staying connected to us through the years, we are lucky to know both of you. To **Arty** (and **Gerard**), the two welcoming faces of the *Sweet Corner* in Den Helder, everything you make is phenomenal but there is a special place in my heart for your wonton soup and iced coffee – true comfort food.

A big thank you to my **family** and **friends** around the globe who have always believed in me. To all my **aunties**, **uncles**, **cousins**, and **Archie** and **Granny** – whether you visited me, checked-in, or just sent your love, you had a part to play in this work. A special thank you to my family already in the Netherlands: **Aunty Shaki** and my cousins **Sabrina** and **Joshua**. Having you here has made it much easier, and although lock downs, health risks, and pregnancies have limited how much time we've been able to spend together, every moment with you all has refilled my cup.

Thank you to **Nishantha**, **Suresh**, **Saman**, **Podi** (Sampath), and the rest at the **Foundation of Goodness** who have been cheering me on for over a decade now.

To **Rachel** (and **Thierry**). Rach, here I am finishing another degree – you've seen me through all three. Your love and support have meant the world to me, and I'd be lost without it. All it takes is a quick chat and you fill me up when I am empty and remind me of how beautiful everything is. You've been there for nearly all the phases of my academic career and have been on board with every part of it. You've pulled me up when I could barely figure myself out and helped reframe my perspective so that I could see my own potential and find a way forward. Every time we talk, I feel full of joy, motivation, and peace, and it's those moments that have helped push me along this journey. I am so excited for what comes next for both of us.

To my in-laws, **Diane** and **Dave**, thank you for always believing that I could do this and for supporting us in so many ways over the years. An especially big thank you to Diane who lived with us for a few months to take care of Bodhi while I wrote a paper.

A very special thank you to **my mother**, who has stuck with me through all my academic adventures and embraced them with encouragement and confidence. I am eternally grateful for all the experiences we've shared over the years and all the ways in which we've grown. You have been a rock for me as I've navigated academia and have shown genuine interest in my work and struggles, even if it can sometimes be very complicated.

The biggest thank you of all to my boys. My house may always be messy, but it's ok because I have you: **Ollie** and **Reuben**, I never thought in a million years that I'd have cats (thank you Nina, this is your doing). You are the sweetest not-so-little boys. Watching you two grow up and play together (and with Bodhi) has been such a joy. You've been there for the ups and downs and have both constantly showered me with cuddles and love. Occasionally you break my stuff and interrupt my work, but it's ok. There's nothing like finally sitting on the couch after a long day knowing you'll come sit with me, Ollie.

Thank you, **Scott**, my life partner for weathering this storm with me from the very beginning. You've pushed me through two degrees, and that is no small feat. You have been in the thick of it and stood by me through and through. There are so many things that I could list here, but I'll just say that this would never have seen the light of day without you. Period. Your patience, love, understanding, encouragement, and unwavering support have made all the difference. You believed in me when I didn't, lifted me up when I couldn't, and lightened the load when I needed a little extra support. You inspire me to be a better person and better academic and to use these skills to make the world a better place. These five years have truly been an adventure, we got married, moved to a new country, bought a house, got some cats, became parents, and are now seeing the turning of this page. I cannot thank you enough for sticking by me through this, for your amazing cooking, and for letting me be a theory-crafting backseat-gamer while you play Elden Ring. It's been tough, but I am so thankful for this little slice of life we've built here in Noord Holland, I am lucky to be sharing it with you.

And to my sweetest little love, **Bodhi**. I couldn't have anticipated it, but in many ways, I became a better scientist since you arrived. I am more intentional, understanding my limitations better as to value my time and energy and prioritize my own well-being to ensure that I am happy and ok for you. You have slowed me down and taught me to appreciate the simple moments in between the intensity of research. You are my biggest source of motivation, and I hope you can look back on this and know that you made me a better person. You are named after a tree, Sri Lanka's Jaya Sri Maha *Bodhi* Tree, which is the oldest known living human-planted tree in the world. Fittingly, the Bodhi tree is an ancient symbol of transformation and our inward journey, which has been one of the biggest parts of this PhD experience for me – challenging but rewarding. I am overjoyed to share this moment with you.

CURRICULUM VITAE

Tara Avanthi Mahendrarajah was born on the 12th of September 1989 in Massachusetts, United States. In 2012, she completed her Bachelor of Science degree in Microbiology (*summa cum laude*) at the University of Massachusetts Amherst. After graduating, Tara volunteered at a rural empowerment program, The Foundation of Goodness in Seenigama, Sri Lanka. There she taught computer literacy, English, assisted in the preschool, and worked in the women's empowerment center. She spent the next few years working in various educational settings until moving to Los Angeles in 2016 to work on a master's degree at California State University Northridge (CSU Northridge) under the guidance of Dr. Rachel Mackelprang. Tara's research focused on assessing the carbon-processing strategies of ancient permafrost microbial communities by linking enzyme activity to metabolic potential predicted using metagenomics. During her time at CSU Northridge, Tara was an instructor of introductory microbiology lab courses for undergraduate students interested in clinical healthcare fields and served as the President of the Microbiology Student Association. Tara obtained her Master of Science in Biology (*with Distinction*) in 2018 and moved to the Netherlands in early 2019 to begin her PhD research at the Royal Netherlands Institute for Sea Research (NIOZ) supervised by Dr. Anja Spang. Transitioning away from microbial ecology in terrestrial settings, Tara's PhD work focused primarily on microbial and molecular evolution using various phylogenetic and comparative genomics approaches, with a goal of understanding the earliest periods of cellular evolution – the results of which are discussed in this thesis. Currently, Tara is a postdoctoral researcher continuing her work on the tree of life with a focus on metabolic inventions across a unique group of Archaea and examining the gene contributions from the prokaryotic ancestors involved in eukaryogenesis.

LIST OF PUBLICATIONS

For the most recent list of publications, please see: https://orcid.org/0000-0001-7032-6581

Mahendrarajah, T.A. (2024) Shedding light on life's deep evolutionary history. In Natuurkundige Voordrachten nieuwe reeks no. 102, Kon. Mij. (pp. 87-96). Voor Natuurkunde 'Diligentia', Den Haag, Netherlands (*in press*)

Moody, E. R. R., Álvarez-Carretero, S., **Mahendrarajah, T. A.**, Clark, J. W., Betts, H. C., Dombrowski, N., Szánthó, L. L., Boyle, R. A., Daines, S., Chen, X., Lane, N., Yang, Z., Shields, G. A., Szöllősi, G. J., Spang, A., Pisani, D., Williams, T. A., Lenton, T. M., & Donoghue, P. C. J. (2024). The nature of the last universal common ancestor and its impact on the early Earth system. *Nature Ecology & Evolution*.

Mahendrarajah, T. A., Moody, E. R. R., Schrempf, D., Szánthó, L. L., Dombrowski, N., Davín, A. A., Pisani, D., Donoghue, P. C. J., Szöllősi, G. J., Williams, T. A., & Spang, A. (2023). ATP synthase evolution on a cross-braced dated tree of life. *Nature Communications*, 14(1), 7456. (**In this thesis: Chapter 3**)

Spang, A., **Mahendrarajah, T. A.**, Offre, P., & Stairs, C. W. (2022). Evolving Perspective on the Origin and Diversification of Cellular Life and the Virosphere. *Genome Biology and Evolution*, 14(6). (**In this thesis: Chapter 5**)

Moody, E. R. R., **Mahendrarajah, T. A.**, Dombrowski, N., Clark, J. W., Petitjean, C., Offre, P., Szöllősi, G. J., Spang, A., & Williams, T. A. (2022). An estimate of the deepest branches of the tree of life from ancient vertically evolving genes. *ELife*, 11. (**In this thesis: Chapter 2**)

Coleman, G. A., Davín, A. A., **Mahendrarajah, T. A.**, Szánthó, L. L., Spang, A., Hugenholtz, P., Szöllősi, G. J., & Williams, T. A. (2021). A rooted phylogeny resolves early bacterial evolution. *Science*, 372(6542). (**In this thesis: Chapter 4**)

Dombrowski, N., **Mahendrarajah, T.**, Gross, S. T., Eme, L., & Spang, A. (2021). Archaea. In Practical Handbook of Microbiology (pp. 229–248). *CRC Press*. (CC BY-NC-ND 4.0). doi:10.1201/9781003099277-23. (**In this thesis: Chapter 6**)

Mackelprang, R., Burkert, A., Haw, M., **Mahendrarajah, T.**, Conaway, C. H., Douglas, T. A., & Waldrop, M. P. (2017). Microbial survival strategies in ancient permafrost: insights from metagenomics. *The ISME Journal*, 11(10), 2305–2318.

I tell my students, when you get these jobs that you have been so brilliantly trained for, just remember that your real job is that if you are free, you need to free somebody else. If you have some power, then your job is to empower somebody else.

- Toni Morrison