RESEARCH ARTICLE SUMMARY

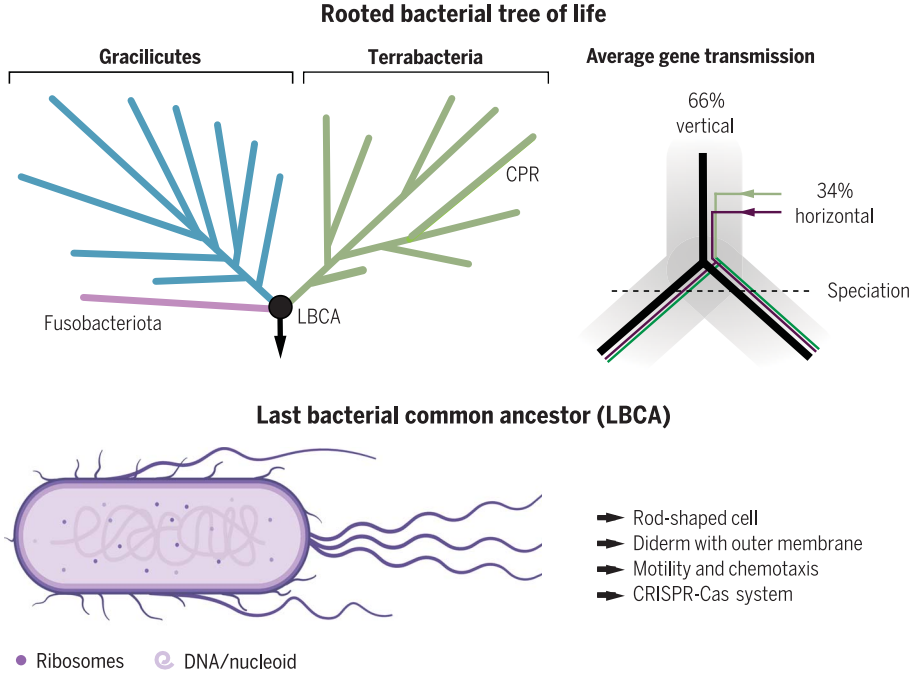
BACTERIAL PHYLOGENY

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‡*

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.



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.

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. ■

The list of author affiliations is available in the full article online.
†These authors contributed equally to this work.
‡These authors contributed equally to this work.
*Corresponding author. Email: p.hugenholtz@uq.edu.au (P.H.); ssolo@elte.hu (G.J.Sz.); tom.a.williams@bristol.ac.uk (T.A.W.)
Cite this article as G. A. Coleman *et al.*, *Science* **372**, eabe0511 (2021). DOI: 10.1126/science.abe0511

RESEARCH ARTICLE

BACTERIAL PHYLOGENY

A rooted phylogeny resolves early bacterial evolution

Gareth A. Coleman[1]†, Adrián A. Davín[2]†, Tara A. Mahendrarajah[3], Lénárd L. Szánthó[4,5], Anja Spang[3,6], Philip Hugenholtz[2]‡*, Gergely J. Szöllősi[4,5,7]‡*, Tom A. Williams[1]‡*

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.

[1]School of Biological Sciences, University of Bristol, Bristol BS8 1TQ, UK. [2]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, Queensland 4072, Australia. [3]Department of Marine Microbiology and Biogeochemistry, NIOZ, Royal Netherlands Institute for Sea Research, 1790 AB Den Burg, Netherlands. [4]Department of Biological Physics, Eötvös Loránd University, 1117 Budapest, Hungary. [5]MTA-ELTE "Lendület" Evolutionary Genomics Research Group, 1117 Budapest, Hungary. [6]Department of Cell- and Molecular Biology, Uppsala University, SE-75123 Uppsala, Sweden. [7]Institute of Evolution, Centre for Ecological Research, 1121 Budapest, Hungary.
†These authors contributed equally to this work.
‡These authors contributed equally to this work.
***Corresponding author. Email: p.hugenholtz@uq.edu.au (P.H.); ssolo@elte.hu (G.J.Sz.); tom.a.williams@bristol.ac.uk (T.A.W.)**

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.

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.

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 non-vertical 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

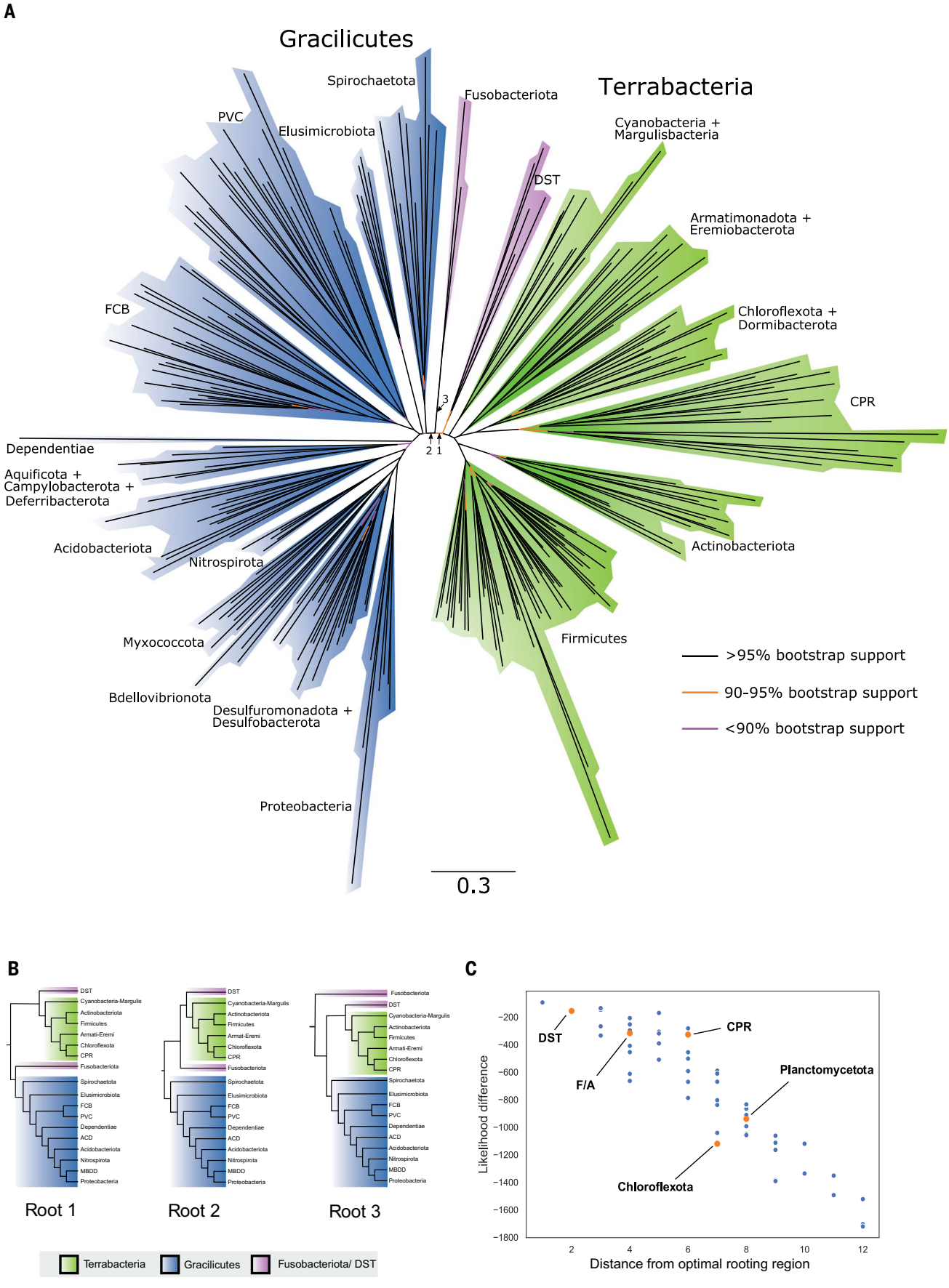


Fig. 1. A rooted phylogeny of Bacteria. (**A**) We used gene tree-species tree reconciliation to infer the root of the bacterial tree. The unrooted maximum likelihood phylogeny was inferred from a concatenation of 62 marker genes under the best-fitting model, LG+C60+R8+F, which accounts for site heterogeneity 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.

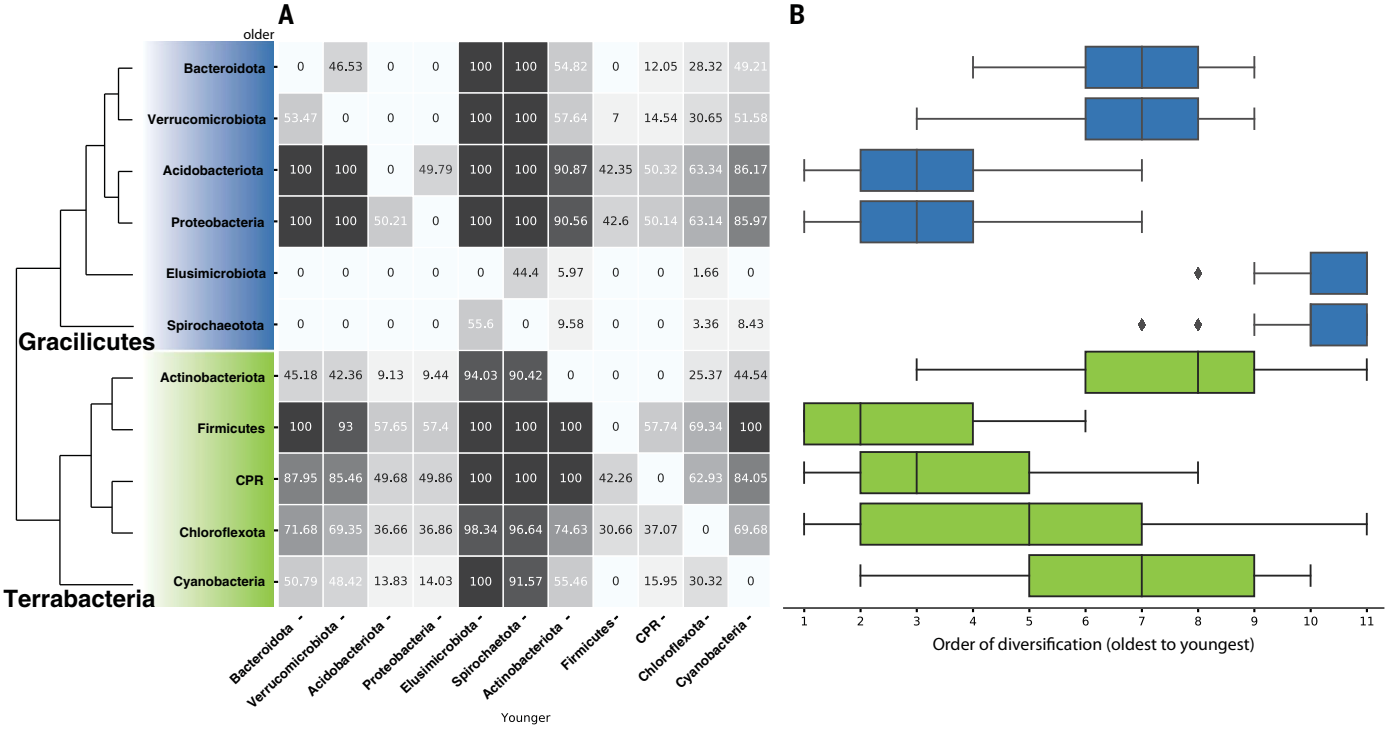


Fig. 2. Relative crown group ages of major bacterial phyla. Gene transfers that occurred during the diversification of Bacteria provide a record of the temporal sequence of events. We used the information provided by directional (donor-to-recipient) patterns of gene transfer to infer the relative ages 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.

mean verticality estimated to be 64% in the focal and 68% in the secondary dataset.

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

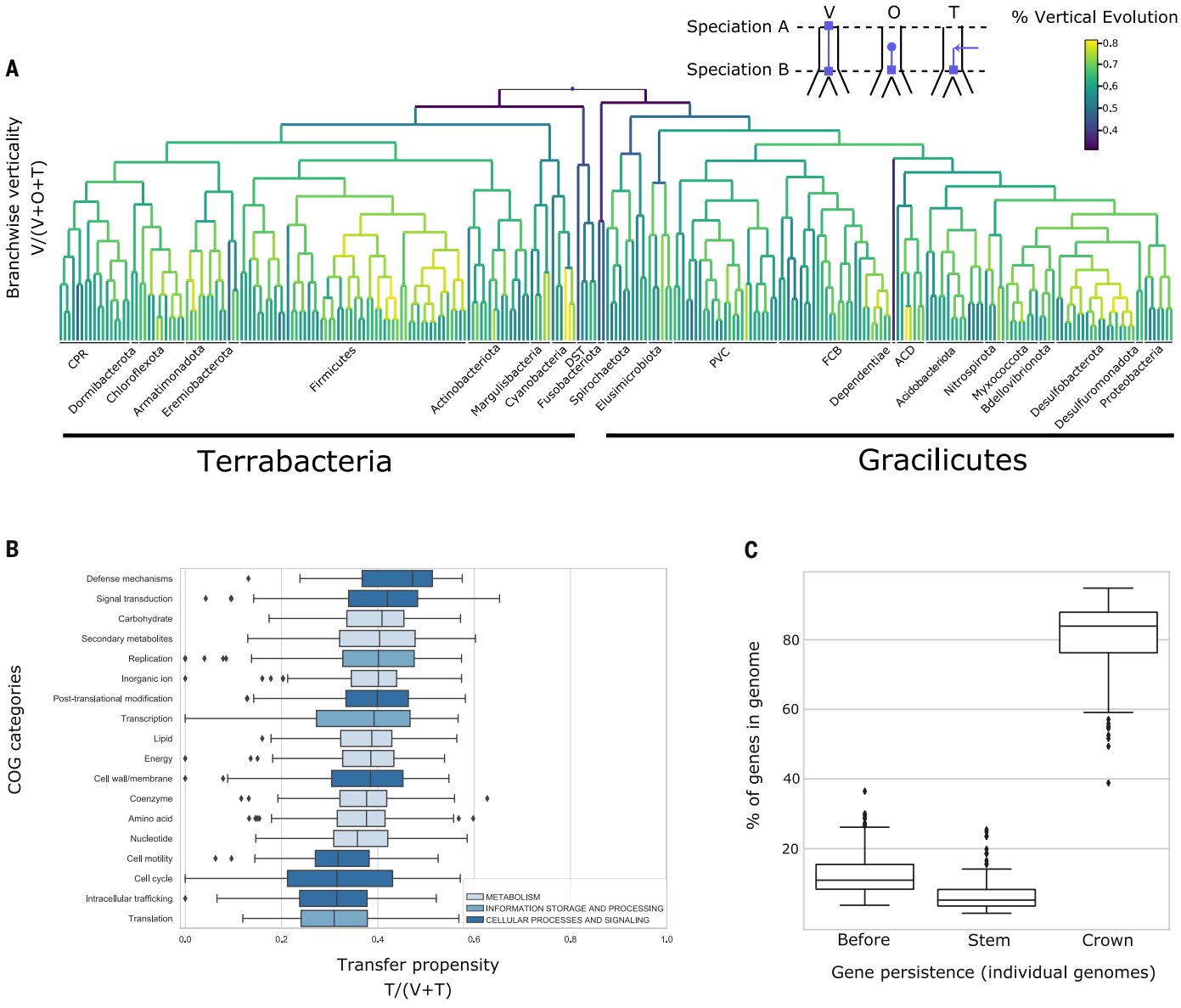


Fig. 3. The verticality of bacterial genome evolution. (**A**) The rooted bacterial species tree (Fig. 1), with branches colored according to verticality: the fraction of genes at the bottom of a branch that descend vertically from the top of that branch (see inset; V, vertical; O, origination; T, transfer 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.

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.

Moderate support for the presence of the shape-determining proteins MreB (PP = 0.9, 0.88, and 0.73 for roots 1 to 3, respectively, as depicted in Fig. 1B), MreC (PP = 0.82/0.79/0.57), and MreD (PP = 0.86/0.83/0.63) 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

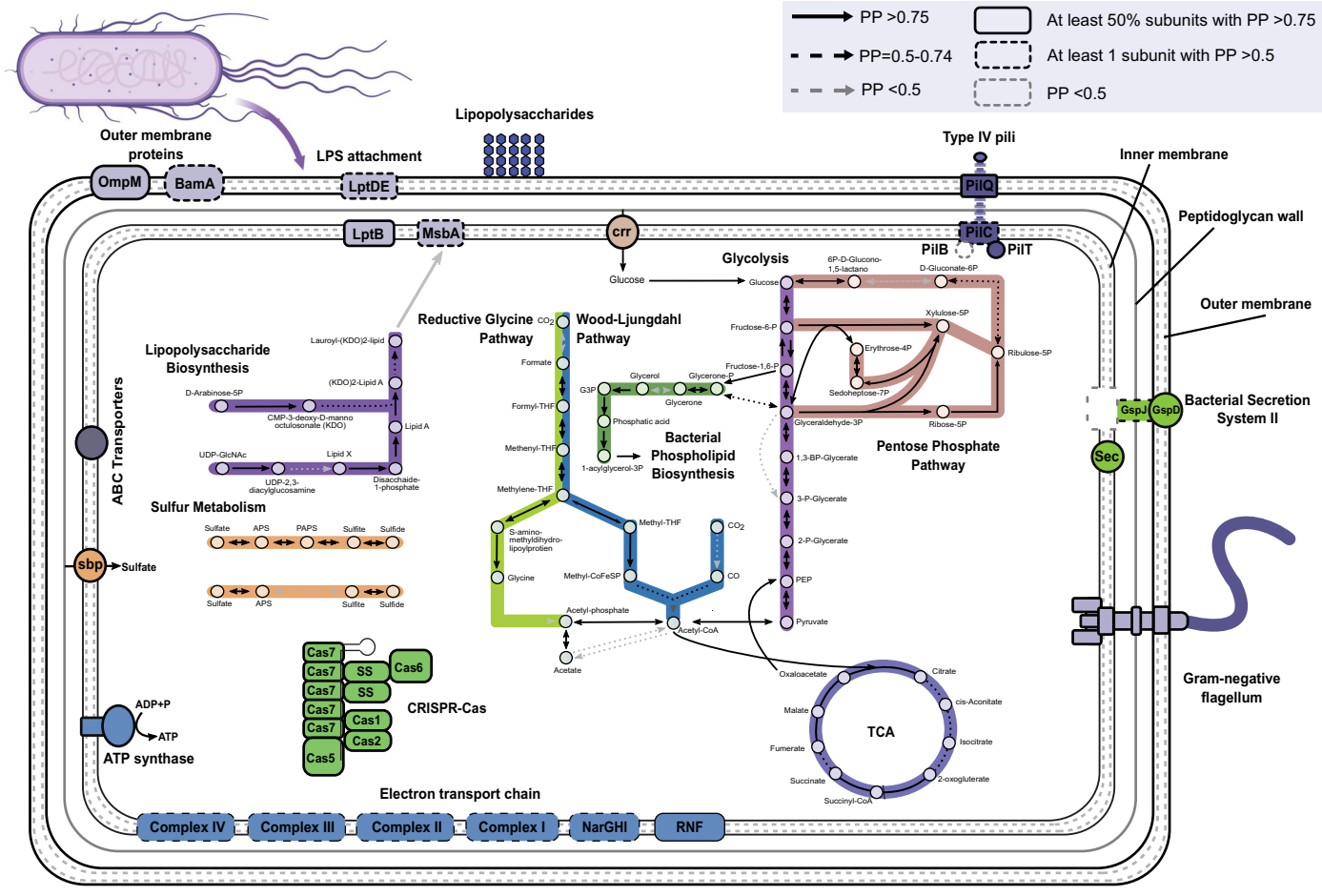


Fig. 4. Ancestral reconstruction of the last bacterial common ancestor (LBCA). The reconstruction is based on genes that could be mapped to at least one branch within the root region with a PP > 0.5 (figs. S17 and S18) (*36*). The presence of a gene within a pathway is indicated as shown in the key. 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).

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*).

REFERENCES AND NOTES

1. T. A. Williams *et al.*, Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E4602–E4611 (2017). doi: 10.1073/pnas.1618463114; pmid: 28533395
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). doi: 10.1073/pnas.92.7.2441; pmid: 7708661
3. J. P. Gogarten *et al.*, Evolution of the vacuolar H+-ATPase: Implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.* **86**, 6661–6665 (1989). doi: 10.1073/pnas.86.17.6661; pmid: 2528146
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). doi: 10.1073/pnas.86.23.9355; pmid: 2531898
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). doi: 10.1007/s00709-005-0135-1; pmid: 16389494
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). doi: 10.1093/molbev/msn247; pmid: 18988685
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). doi: 10.1007/s002399910040; pmid: 10795828
8. C. Brochier, H. Philippe, A non-hyperthermophilic ancestor for Bacteria. *Nature* **417**, 244 (2002). doi: 10.1038/417244a; pmid: 12015592
9. T. Cavalier-Smith, Rooting the tree of life by transition analyses. *Biol. Direct* **1**, 19 (2006). doi: 10.1186/1745-6150-1-19; pmid: 16834776
10. J. A. Lake, Evidence for an early prokaryotic endosymbiosis. *Nature* **460**, 967–971 (2009). doi: 10.1038/nature08183; pmid: 19693078
11. L. A. Hug *et al.*, A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016). doi: 10.1038/nmicrobiol.2016.48; pmid: 27572647
12. S. Mukherjee *et al.*, 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat. Biotechnol.* **35**, 676–683 (2017). doi: 10.1038/nbt.3886; pmid: 28604660
13. D. H. Parks *et al.*, A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life.

Nat. Biotechnol. **36**, 996–1004 (2018). doi: 10.1038/nbt.4229; pmid: 30148503
14. D. H. Parks *et al.*, Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017). doi: 10.1038/s41564-017-0012-7; pmid: 28894102
15. C. T. Brown *et al.*, Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015). doi: 10.1038/nature14486; pmid: 26083755
16. Q. Zhu *et al.*, Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477 (2019). doi: 10.1038/s41467-019-13443-4; pmid: 31792218
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). doi: 10.1016/j.cell.2018.02.016; pmid: 29522741
18. C. J. Castelle *et al.*, Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat. Rev. Microbiol.* **16**, 629–645 (2018). doi: 10.1038/s41579-018-0076-2; pmid: 30181663
19. J. P. Beam *et al.*, Ancestral absence of electron transport chains in Patescibacteria and DPANN. *Front. Microbiol.* **11**, 1848 (2020). doi: 10.3389/fmicb.2020.01848; pmid: 33013724
20. C. J. Castelle *et al.*, Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr. Biol.* **25**, 690–701 (2015). doi: 10.1016/j.cub.2015.01.014; pmid: 25702576
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). doi: 10.1038/s41467-019-12171-z; pmid: 31519891
22. A. Graybeal, Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst. Biol.* **47**, 9–17 (1998). doi: 10.1080/106351598260996; pmid: 12064243
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). doi: 10.1080/10635150600697358; pmid: 16861214
24. J. Bergsten, A review of long-branch attraction. *Cladistics* **21**, 163–193 (2005). doi: 10.1111/j.1096-0031.2005.00059.x
25. R. Gouy, D. Baurain, H. Philippe, Rooting the tree of life: The phylogenetic jury is still out. *Philos. Trans. R. Soc. London Ser. B* **370**, 20140329 (2015). doi: 10.1098/rstb.2014.0329; pmid: 26323760
26. J. A. Lake, R. G. Skophammer, C. W. Herbold, J. A. Servin, Genome beginnings: Rooting the tree of life. *Philos. Trans. R. Soc. London Ser. B* **364**, 2177–2185 (2009). doi: 10.1098/rstb.2009.0035; pmid: 19571238
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). doi: 10.1093/molbev/msm096; pmid: 17513883
28. H. Shimodaira, An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508 (2002). doi: 10.1080/10635150290069913; pmid: 12079646
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). doi: 10.1073/pnas.1202997109; pmid: 23043116
30. L. A. David, E. J. Alm, Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* **469**, 93–96 (2011). doi: 10.1038/nature09649; pmid: 21170026
31. 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). doi: 10.1093/sysbio/syt054; pmid: 23925510
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). doi: 10.1093/sysbio/sys026; pmid: 22334342
33. D. M. Emms, S. Kelly, STRIDE: Species Tree Root Inference from Gene Duplication Events. *Mol. Biol. Evol.* **34**, 3267–3278 (2017). doi: 10.1093/molbev/msx259; pmid: 29029342
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). doi: 10.1093/molbev/msz088; pmid: 31004147
35. S. Q. Le, O. Gascuel, An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008). doi: 10.1093/molbev/msn067; pmid: 18367465
36. See 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). doi: 10.1080/10635150490522304; pmid: 15545256
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). doi: 10.1093/nar/30.7.1575; pmid: 11917018
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). doi: 10.1093/bib/bbx117; pmid: 28968633
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). doi: 10.1073/pnas.1420858112; pmid: 25964353
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). doi: 10.1073/pnas.1716667115; pmid: 29358391
42. A. A. Davín *et al.*, Gene transfers can date the tree of life. *Nat. Ecol. Evol.* **2**, 904–909 (2018). doi: 10.1038/s41559-018-0525-3; pmid: 29610471
43. C. Chauve *et al.*, MaxTiC: fast ranking of a phylogenetic tree by maximum time consistency with lateral gene transfers. bioRxiv 127548 [Preprint]. 6 October 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). doi: 10.1073/pnas.0606318104; pmid: 17213324
45. W. F. Doolittle, Phylogenetic classification and the universal tree. *Science* **284**, 2124–2128 (1999). doi: 10.1126/science.284.5423.2124; pmid: 10381871
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). doi: 10.1073/pnas.0610699104; pmid: 17261804
47. T. Dagan, W. Martin, The tree of one percent. *Genome Biol.* **7**, 118 (2006). doi: 10.1186/gb-2006-7-10-118; pmid: 17081279
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–E1603 (2013). doi: 10.1073/pnas.1211371110; pmid: 23576716
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). doi: 10.1073/pnas.96.7.3801; pmid: 10097118
50. P. Puigbò, Y. I. Wolf, E. V. Koonin, The tree and net components of prokaryote evolution. *Genome Biol. Evol.* **2**, 745–756 (2010). doi: 10.1093/gbe/evq062; pmid: 20889655
51. R. Liu, H. Ochman, Stepwise formation of the bacterial flagellar system. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7116–7121 (2007). doi: 10.1073/pnas.0700266104; pmid: 17438286
52. F. El Baidouri, C. Venditti, S. Suzuki, A. Meade, S. Humphries, Phenotypic reconstruction of the last universal common ancestor reveals a complex cell. bioRxiv 2020.08.20.260398 [Preprint]. 21 August 2020. https://doi.org/10.1101/2020.08.20.26039.
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). doi: 10.1073/pnas.1614083113; pmid: 27702904
54. L. C. Antunes *et al.*, Phylogenomic analysis supports the ancestral presence of LPS-outer membranes in the Firmicutes. *eLife* **5**, e14589 (2016). doi: 10.7554/eLife.14589; pmid: 27580370
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). doi: 10.1111/mmi.14469; pmid: 31975449
56. N. Taib *et al.*, Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. *Nat. Ecol. Evol.* **4**, 1661–1672 (2020). doi: 10.1038/s41559-020-01299-7; pmid: 33077930
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). doi: 10.1038/nrmicro.2016.85; pmid: 28232669
58. I. Sánchez-Andrea *et al.*, The reductive glycine pathway allows autotrophic growth of *Desulfovibrio desulfuricans*. *Nat. Commun.* **11**, 5090 (2020). doi: 10.1038/s41467-020-18906-7; pmid: 33037220
59. M. C. Weiss *et al.*, The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.* **1**, 16116 (2016). doi: 10.1038/nmicrobiol.2016.116; pmid: 27562259
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 Bioenerg.* **1837**, 964–981 (2014). doi: 10.1016/j.bbabio.2014.02.001; pmid: 24513196
61. F. L. Sousa, S. Nelson-Sathi, W. F. Martin, One step beyond a ribosome: The ancient anaerobic core. *Biochim. Biophys. Acta Bioenerg.* **1857**, 1027–1038 (2016). doi: 10.1016/j.bbabio.2016.04.284; pmid: 27150504
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). doi: 10.1093/gbe/evw114; pmid: 27189979
63. G. Fuchs, Alternative pathways of carbon dioxide fixation: Insights into the early evolution of life? *Annu. Rev. Microbiol.* **65**, 631–658 (2011). doi: 10.1146/annurev-micro-090110-102801; pmid: 21740227
64. T. Nunoura *et al.*, A primordial and reversible TCA cycle in a facultatively chemolithoautotrophic thermophile. *Science* **359**, 559–563 (2018). doi: 10.1126/science.aao3407; pmid: 29420286
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). doi: 10.1038/nrmicro3365; pmid: 25383604
66. K. S. Makarova *et al.*, Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol.* **9**, 467–477 (2011). doi: 10.1038/nrmicro2577; pmid: 21552286
67. E. V. Koonin, K. S. Makarova, Origins and evolution of CRISPR-Cas systems. *Philos. Trans. R. Soc. London Ser. B* **374**, 20180087 (2019). doi: 10.1098/rstb.2018.0087; pmid: 30905284
68. J. K. Nuñez *et al.*, Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat. Struct. Mol. Biol.* **21**, 528–534 (2014). doi: 10.1038/nsmb.2820; pmid: 24793649
69. K. S. Makarova *et al.*, An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **13**, 722–736 (2015). doi: 10.1038/nrmicro3569; pmid: 26411297
70. R. Barrangou *et al.*, CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007). doi: 10.1126/science.1138140; pmid: 17379808
71. E. V. Koonin, The origins of cellular life. *Antonie van Leeuwenhoek* **106**, 27–41 (2014). doi: 10.1007/s10482-014-0169-5; pmid: 24756907
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). doi: 10.1038/s41579-019-0205-6; pmid: 31142823
73. A. C. J. Roth, G. H. Gonnet, C. Dessimoz, Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* **9**, 518 (2008). doi: 10.1186/1471-2105-9-518; pmid: 19055798
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). doi: 10.1093/molbev/mst010; pmid: 23329690
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). doi: 10.1186/1471-2148-10-210; pmid: 20626897
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). doi: 10.1093/molbev/msu300; pmid: 25371430
77. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015). doi: 10.1038/nmeth.3176; pmid: 25402007
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). doi: 10.1093/bioinformatics/btz710; pmid: 31566657
79. J. Huerta-Cepas *et al.*, Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper.

Mol. Biol. Evol. **34**, 2115–2122 (2017). doi: 10.1093/molbev/msx148; pmid: 28460117
80. G. Coleman *et al.*, Extended Data for A rooted phylogeny resolves early bacterial evolution, Version 9, Figshare (2020); https://doi.org/10.6084/m9.figshare.12651074.v9.

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 MATERIALS

science.sciencemag.org/content/372/6542/eabe0511/suppl/DC1
Materials and Methods
Figs. S1 to S18
Tables S1 to S13
References (*81*–*95*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

28 July 2020; resubmitted 5 November 2020
Accepted 1 April 2021
10.1126/science.abe0511

Science

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. Szöllosi and Tom A. Williams

Science **372** (6542), eabe0511.
DOI: 10.1126/science.abe0511

Reconstructing ancestral bacteria

The origin of the eubacteria and phylogenetic relationships between subgroups have been difficult to resolve. Applying a phylogenetic analysis and recent computational methods to the expanded diversity of bacterial sequences from metagenomic analyses, Coleman *et al.* infer the root of the eubacterial tree (see the Perspective by Katz). The root was determined without using the Archaea as an outgroup, to avoid the possibility of a false result due to long branch attraction. This method places the eubacterial root in the neighborhood of Fusobacteriota. Using this information, the authors reconstructed the eubacterial ancestor, identifying that this organism likely had a double-membrane cell envelope, flagellum-mediated motility, antiphage defense mechanisms, and diverse metabolic pathways.

Science, this issue p. eabe0511; see also p. 574

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.