

Phylogenetic reconciliation supports a methanogenic ancestor of the Archaea and a derived origin for host-associated lineages

Received: 10 June 2026

Accepted: 17 July 2026

Published online: 13 August 2026

 Check for updates

Wen-Cong Huang (黄文聪)^{1,2}, Nina Dombrowski^{1,3}, Tara A. Mahendrarajah¹, Noah Wahl⁴, Alexandros Stamatakis^{4,5,6}, Gergely J. Szöllösi^{7,8}✉, Tom A. Williams⁹✉ & Anja Spang^{1,2}✉

The phylogeny of the Archaea continues to be revisited and revised as new groups are discovered and phylogenetic methods improve, but key questions about their early evolution remain unanswered. It has been suggested that the root of the Archaea may lie on, or potentially within, any of three major groups - the Euryarchaeota, TACK+Asgard clade, and DPANN, the last of which includes many host-associated and genome-reduced lineages. These root hypotheses make starkly different predictions about the nature of early archaeal evolution: for example, a root on or within DPANN might suggest a small-genome and host-associated ancestor, with methanogenesis, the hall-mark metabolism of the Archaea, evolving later. Here, we investigate the position of the archaeal root and the nature of the last archaeal common ancestor using a range of phylogenetic approaches, including the best available site- and branch-heterogeneous substitution models, and new gene tree-species tree reconciliation models that capture changes in rates of gene duplication, loss and transfer across the phylogeny. Our analyses converge on a narrow archaeal root region at or near the base of the Euryarchaeota, supporting hypotheses in which the Last Archaeal Common Ancestor (LACA) was a complex, free-living (hyper-)thermophilic methanogen. We recover DPANN as the sister group to TACK and Asgard archaea, and suggest that their genome evolution has been characterised by episodes of genome streamlining and expansion, driven by gene loss and transfer.

Over the past 15 years, developments in environmental genomics have led to the discovery of major new lineages that have fuelled new interest in, and debates about, the deep structure of the tree of life^{1–7}. In particular, the discovery of clades comprising genetically diverse, small-genome-sized, host-associated archaea and bacteria—i.e., the “DPANN” archaea (an acronym of the first described phyla: Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaeota, now referred to as Nanobdellati⁸ and hereafter informally as DPANN) and the “CPR” bacteria (Candidate Phyla Radiation bacteria,

now termed Patescibacteria⁹/Minisyncocota¹⁰ and hereafter informally referred to as CPR)—has catalysed new controversies about the antiquity of symbiosis and parasitism in life’s history. Initial phylogenetic analyses placed DPANN and CPR at the base of their respective domains⁴, potentially implying archaeal and bacterial ancestors of only moderate size and cellular and metabolic complexity. Subsequent analyses using improved phylogenetic methods resolved CPR as a derived clade related to Chloroflexota^{11–15} and suggested that the last bacterial common ancestor was likely a complex and free-living cell

A full list of affiliations appears at the end of the paper. ✉ e-mail: gergely.szollosi@oist.jp; tw2316@bath.ac.uk; anja.spang@nioz.nl

akin in complexity to modern bacteria¹². However, resolving or approaching a similar degree of consensus on the position of DPANN within Archaea has proven more challenging, with published analyses recovering phylogenies that have starkly different consequences for our understanding of archaeal origins and the sequence of metabolic evolution within this domain^{16–20}.

The primary question is whether DPANN branch at the base of the archaeal domain, or alternatively whether DPANN occupy a derived position, as sister to or branching within one of the other major lineages, which include “Euryarchaeota” (also referred to as Methanobacteriati⁸ and hereafter informally as Euryarchaeota), Thermoproteota (previously referred to as “TACK” Archaea²¹, now as Thermoproteati⁸ and hereafter as TACK) and “Asgard” archaea (Asgardarchaeota²²/ Promethearchaeota²³; hereafter referred to as Asgard)^{4,14,17–20,24}. Resolving the placement of DPANN in the archaeal tree is critically important because it informs the nature of LACA and the tempo and mode of archaeal genome evolution. This is because DPANN predominantly, though not exclusively^{24–28}, comprise organisms with reduced genomes and apparently limited metabolic capabilities. Indeed, the first cultivated representatives^{29–36} of DPANN were shown to be obligate archaeal ectosymbionts depending on distinct archaeal^{24–26,37} and potentially even bacterial hosts³⁸. If the DPANN branch at the base of the Archaea, this could suggest that LACA was a relatively simple and small-genome-sized organism, with methanogenesis—the hallmark metabolism of the Archaea^{39–41}—evolving later, after the divergence of DPANN from Euryarchaeota, TACK and Asgard archaea. Alternatively, if the archaeal root lies within one of those alternative lineages with DPANN being derived, LACA may instead have been a bona fide complex, free-living organism.

Specifically, determining the position of DPANN within the archaeal phylogeny has proven challenging. Some recent studies have resolved DPANN as the earliest-diverging branch in the archaeal tree^{13,14,18}, or with, or within, Euryarchaeota¹⁷, while others have suggested that DPANN might actually be polyphyletic, evolving multiple times from within the Euryarchaeota via parallel events of genome reduction^{42–44}. Each of these inferred positions has been suggested to be induced by various phylogenetic artefacts. For example, the distinctive evolutionary mode of DPANN—which, as symbionts, may both evolve fast and lose many genes otherwise conserved in archaea—might artifactually draw them to the root of the tree due to long branch attraction (LBA)^{17,19,20,45}. On the other hand, alternative DPANN placements at derived positions within the archaeal tree have also been suggested to result from phylogenetic artefacts. For example, the placement of DPANN with Euryarchaeota in some analyses has been suggested to result from unmodelled horizontal gene transfer (HGT) between DPANN symbionts and their euryarchaeotal hosts^{46,47}, or from systematic phylogenetic error⁴⁸. A recent analysis using a newly-developed non-stationary phylogenetic model that captures variation in sequence composition across both the sites of the alignment and the branches of the tree—two pervasive but often un-modelled features of sequence evolution—provided support for a derived position of DPANN branching sister to Euryarchaeota, with the root of the archaeal tree lying between the TACK+Asgard archaea and all other archaea¹⁷.

The studies summarised above used a range of distinct approaches to root the archaeal phylogeny. The most widely used rooting approach is to use an outgroup—that is, to include a bacterial outgroup to root the archaeal tree on the inter-domain branch. This approach is conceptually straightforward but may be problematic for rooting entire domains, because the long inter-domain branch might induce errors in the within-Archaea topology^{49,50}. In particular, the basal position of DPANN in published trees might result from long-branch attraction between the DPANN and inter-domain stems. As an alternative to outgroup rooting, non-reversible or non-stationary models can be used to root the tree based on capturing directional aspects of the evolutionary process, such as long-term trends in amino

acid substitution patterns or changing amino acid compositions over time^{51–55}. However, the power of these approaches to root trees depends on the strength of the time-dependent signal in the data and the ability of the model to adequately describe the underlying processes⁵³. These signals are sometimes weak for anciently diverged sequences⁵¹. A third approach is to use phylogenetic reconciliation, which can assess distinctly rooted species trees based on inferred evolutionary events that have a clear direction in time, such as gene duplications, losses and transfers^{12,18,56,57}. However, as with substitution models, the accuracy of the results depends on the adequacy with which the model captures the underlying processes¹². In addition to these modelling concerns, different studies have used distinct sets of phylogenetic “marker” genes—genes that are thought to provide a reliable trace of vertical archaeal evolution—and protein families in the case of reconciliations with the differing properties of these datasets introducing additional variation into the results^{13,17–20,45,48}. While none of the available rooting methods or datasets are conclusive, the hope is that, as methods improve, they may converge on a consistent root position supported—or at least not strongly rejected—by distinct analytical approaches.

Here, we investigate the root of the archaeal tree, using the best available outgroup-rooting methods and recently developed non-reversible and non-stationary substitution models^{51,58,59}. For the first time, we also apply a new class of more flexible and better-fitting phylogenetic reconciliation methods that we recently developed⁵⁶, and which allow the model of genome evolution to vary across the rooted species tree. The reconciliation method not only allows us to assess the support for various root hypotheses but also to reconstruct historical patterns of genome evolution across Archaea.

Results

Outgroup rooting has only limited power to resolve the root of Archaea

We re-visited suitable markers and outgroup rooting on an archaeal tree using a focal dataset of 257 Archaea and 103 Bacteria (Supplementary Data 1) based on 25 vertically-evolving markers that did not violate the requirement of domain monophyly (Supplementary Data 2)¹³. Specifically, we applied several rounds of manual curation to assemble a dataset of 257 phylogenetically representative, high-quality (low contamination, high completeness) genomes sampled from across the known diversity of Archaea (Supplementary Data 1) (see “Methods”). The maximum likelihood tree under the best-fitting site-heterogeneous model (LG + C60 + F + R; chosen via the Bayesian Information Criterion (BIC)) placed the archaeal root between DPANN and all other archaea (Fig. S1, DPANN root, 93.3 and 61.18 ultrafast bootstrap support⁶⁰, respectively, for the monophyly of DPANN lineage and the monophyly of Euryarchaeota+TACK+Asgard), consistent with some previous analyses^{4,13,18,24,61}. However, previous work has shown that there is sometimes limited information about the archaeal root position in such “tree of life”-scale analyses^{12–14,18,62}, so that the best-known maximum likelihood tree is not necessarily significantly better than trees with alternative root positions. To test this possibility on our data, we generated a range of rooting hypotheses (Supplementary Data 3), including key hypotheses from the literature as well as other plausible root positions (e.g., between each of the major archaeal taxonomic clades), by relocating the bacterial outgroup's attachment point to the archaeal ingroup. For each candidate root, we inferred the best-known ML tree while remaining consistent with that root hypothesis using a constrained tree search in IQ-TREE 2⁶³, and then assessed whether the likelihoods of these trees were significantly worse than the unconstrained ML tree using an approximately unbiased (AU) test⁶⁴. This analysis rejected some hypotheses, such as a root between the Asgard archaea and all other Archaea ($P < 0.05$; Supplementary Data 3). However, various other hypotheses, some of which received literature support, e.g., DPANN root¹⁸, a root between

TACK+Asgard archaea and all other archaea (hereafter, TACK+Asgard root)^{17,45}, a root between a clade comprising Halobacteriota and Thermoplasmatota and all other archaea (hereafter, Halobacteriota+Thermoplasmatota root)^{19,20}, and a root between Euryarchaeota and all other archaea (hereafter, Euryarchaeota root), remained statistically indistinguishable (Supplementary Data 3). These results are also consistent with another recent analysis of the tree of life using site-heterogeneous models⁶², in which the maximum likelihood analysis supported TACK+Asgard on one side of the archaeal root, and Euryarchaeota+DPANN on the other (“Topology 1” in ref. 64); however, this root was statistically indistinguishable from a root between DPANN and all other Archaea based on the AU test.

Inferring an unrooted species tree of Archaea

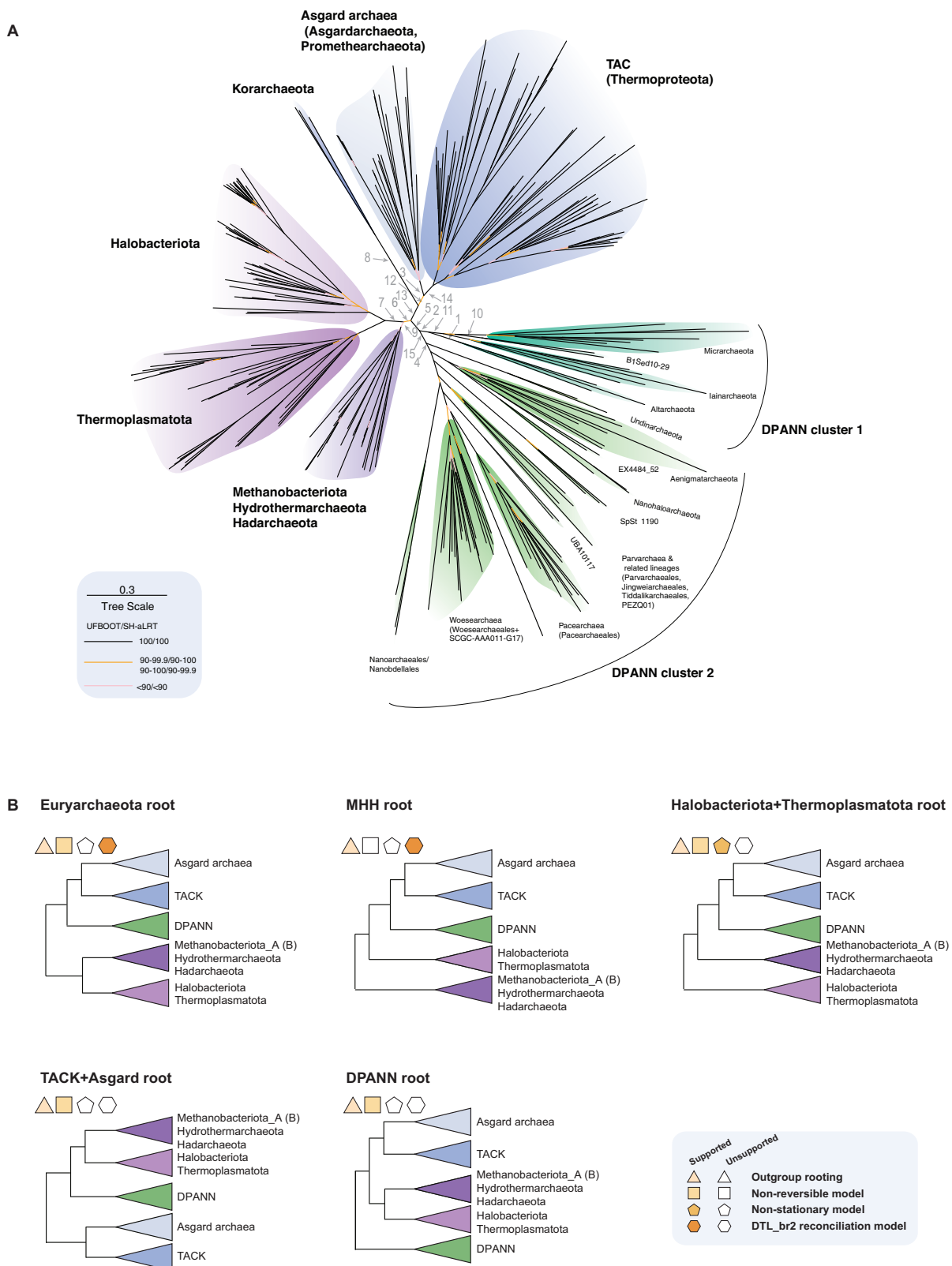
Given the lack of resolution in the outgroup rooting analyses, we next inferred an unrooted phylogenetic tree for subsequent use of outgroup-free rooting using non-stationary and phylogenetic reconciliation models^{17,56–59,65}. Since tree search is computationally prohibitive under these models, we followed a two-step procedure in which we initially estimated an unrooted archaeal species tree using a concatenation of phylogenetic marker genes, then searched for the root position on this fixed tree using a variety of approaches. We extracted the sequences for a set of 51 marker genes previously suggested to evolve vertically within the Archaea⁶¹ from our archaeal focal dataset of 257 representative genomes; manual inspection (see “Methods”) suggested that 45 of those markers were suitable for concatenation (i.e., they had no indication for inter-domain HGT and phylogenetic incongruence) (Supplementary Data 4). We then inferred an unrooted tree using the best-fitting model (LG + C60 + F + R; chosen by BIC; see “Methods”), which accounts for different rates of change among amino acids and variation in evolutionary rate, as well as sequence composition across the sites of the sequence alignment. This phylogeny recovered the clanhood (that is, monophyly conditional on a root outside the group⁶⁶) of three archaeal groups with maximal or near-maximal support based on the ultrafast bootstrap (UFBOOT)⁶⁰ and Shimodaira–Hasegawa approximate likelihood ratio test (SH-aLRT)⁶⁷ metrics: Euryarchaeota (100/99), TACK+Asgard archaea (100/100) and DPANN archaea (100/100) (Fig. 1). To evaluate whether the inferred topology might be affected by across-branch variation in amino acid composition—a feature of sequence evolution not accounted for by the LG + C60 + F + R model—we incrementally filtered out 10%–80% of the most compositionally-biased sites from the alignment and re-inferred the species trees, which produced highly congruent species trees that all recovered the three archaeal groups (Fig. S2). This suggests that key features of our topology are not artefacts of biased sequence composition. Phylogenetic inference using an evenly downsampled, phylum-level dataset produced highly congruent species trees that consistently recovered these three groups and archaeal phylum relationships, indicating that topologies reported here are not affected by taxon sampling or imbalance (Fig. S3). Furthermore, we used Astral-Pro³⁶⁸, a gene tree summary approach, to infer a species tree from all 7059 arCOG (archaeal Cluster of Orthologues) families, including the 45 core genes used for concatenation and 241 additional archaeal gene families found in at least 80% of archaeal genomes (Fig. S4). The inferred topologies were in agreement with previous analyses—for example, recovering Euryarchaeota, TACK+Asgard and DPANN clans, as well as many of the major relationships within each of those groups^{14,17,18,61,62}. Finally, to evaluate the impact of differing evolutionary histories among genes introduced by horizontal gene transfers, we fitted a mixture across sites and trees (MAST)⁶⁹ model to the same concatenation. The majority of sites in the concatenation supported the Euryarchaeota, TACK+Asgard, and DPANN clans and the positions of key archaeal phyla in the tree (Fig. S5 and Supplementary Note 1).

One lineage that has proven difficult to place in previous analyses is the Altiarchaeota, which branch with DPANN but do not appear to be genome-reduced and are predicted to be capable of free-living growth²⁷. We recovered Altiarchaeota as a member of DPANN Cluster 1 (a clade comprising Altiarchaeota, Micrarchaeota, Iainarchaeota and B1Sed10-23)⁶¹ (Fig. 1), in agreement with several previous studies^{61,70–73}. By contrast, other recent studies placed the Altiarchaeota as the earliest-diverging branch within DPANN as a whole^{17,19}. These studies made use of an alternative set of marker genes with a range of functions beyond information processing, including metabolic genes such as subunits of ATP synthase. While an expanded marker set is valuable, a recurrent challenge is the higher rate of horizontal gene transfer in non-ribosomal proteins¹³. We re-analysed these genes and, after filtering out marker genes with instances of HGT within Archaea and between Archaea and Bacteria, recovered Altiarchaeota at the base of Cluster 1 (see Supplementary Note 1 for further details, Fig. S6 and Supplementary Data 5).

Analysis with non-reversible and non-stationary substitution models supports a root within Euryarchaeota

We next used non-reversible models to infer a rooted species tree and a non-stationary substitution model to infer the root position on the fixed unrooted species tree. We first performed analyses under the non-reversible models^{51,74} implemented in IQ-TREE 2⁶³. In these models, exchangeabilities between amino acid states are non-reversible; for example, the probability of change from D (aspartic acid) to E (glutamic acid) is different to the probability from E to D; as a result, trees inferred under the non-reversible models are intrinsically rooted, with the root information derived from long-term trends in protein evolution (a long-term difference in the rate at which exchanges occur between pairs of amino acids; e.g., the relative rate of change of amino acid from D to E was 0.45, while the rate from E to D was 0.19, in our dataset). When applying such considerations to the 257 archaeal species, the 45-marker gene dataset inferred the best known maximum likelihood root between Korarchaeota and all other archaea (Korarchaeota root, rootstrap support: 42.8)⁵¹ under the NONREV + R10 model (Fig. S7A) which provided the best model fit among the models tested (Supplementary Data 6). As in the outgroup-rooting analyses, we used an AU test⁶⁴ on site-wise likelihoods to evaluate whether the Korarchaeota root provided a significantly better fit to the data than alternative positions that are rerooted at every branch of the inferred maximum likelihood tree. As with outgroup rooting, the analysis indicated that a variety of root hypotheses could not be rejected under the NONREV + R10 model for these data, including roots on Korarchaeota, Thermoproteota, DPANN, Euryarchaeota and TACK+Asgard archaea (Supplementary Data 6 and Fig. S7A).

Non-stationary models can also be used to infer the root of a tree, accounting for changes in the substitution process over time. One such model is the GFMix model^{17,58,59}, which models changing nucleotide or amino acid sequence compositions over the branches of the phylogenetic tree. This kind of compositional change is manifest in real data. For example, halophilic archaea were reported to be enriched in D and E while depleted in I (isoleucine) and K (lysine) compared to non-halophilic archaea^{48,75}. GFMix is particularly appealing for analyses of deep microbial evolution because it combines the modelling of across-site compositional heterogeneity (as in the site-heterogeneous C60 model described above) with across-branch compositional variation—both pervasive features of these datasets—in a single analysis. When applied to our 45-gene supermatrix, a root within the Euryarchaeota, i.e., the Halobacteriota+Thermoplasmatota root^{19,20} had the highest likelihood with all other tested roots being rejected by an AU test ($P < 0.05$, Supplementary Data 7). By contrast, a recent analysis of a deep archaeal phylogeny using GFMix and a different marker gene set, tested a range of root hypotheses and reported that a root between TACK+Asgard and all other Archaea had the highest likelihood among



the trees tested, with DPANN being derived from within a paraphyletic Euryarchaeota¹⁷. We repeated this analysis on the original dataset, but including a broader set of root hypotheses (9 more positions in addition to the DPANN root and TACK+Asgard root) (Supplementary Data 8) and found that several of these had significantly higher likelihoods. Among those we tested, the Halobacteriota+Thermoplasmatota root^{19,20} had the highest likelihood in agreement with our

new results from the 45-gene concatenate (being 8.6 likelihood units higher than the TACK+Asgard root; see Supplementary Data 7). All other tested roots, including a root between TACK+Asgard and all other Archaea¹⁷, were rejected by the AU test ($P < 0.05$; see Supplementary Data 8).

Overall, our analyses of these datasets using outgroup rooting, non-reversible, and non-stationary models suggest that the root of the

Fig. 1 | An unrooted phylogeny of the archaeal tree of life and depictions of alternative root hypotheses. A An unrooted maximum likelihood phylogeny of Archaea, inferred from a supermatrix of 45 vertically-evolving marker genes under the best-fitting site-heterogeneous model, LG + C60 + F + R. The numbers along the branches of the tree indicate the various root positions tested. **B** Schematic representation of the most supported root positions using different root inference methods, as indicated by the colour-coded shapes in the figure inlet: outgroup rooting, non-reversible, non-stationary (GFMix) and the DTL_br2 reconciliation models (See Methods) (i.e., that cannot be rejected by the approximately unbiased

(AU) test; $P > 0.05$, the test is a one-sided statistical test). AU test results of outgroup rooting, non-reversible model, non-stationary (GFMix) model, the DTL_br2 reconciliation model, and the best-fitting reconciliation model are provided in Supplementary Data 3, 6, 7, 12 and 13. Among these, the DTL_br2 reconciliation model supports a root between Euryarchaeota and all other Archaea and a root between a clade comprising Methanobacteriota_A, Methanobacteriota_B, Hydrothermarchaeota and Hadarchaeota on one side and all other Archaea on the other (MHH root). The best-fitting reconciliation model supports a root between Euryarchaeota and all other Archaea. Scale Bar: average substitutions per site.

Archaea remains a challenging problem, yielding rather similar likelihoods for a variety of plausible root hypotheses, most of which are within 3 nodes of one another on the unrooted species tree. To complement these analyses, we explored rooting using gene tree-species tree reconciliation, which makes use of a substantially larger fraction of the available genome data and might therefore have greater resolving power to assess contrasting root hypotheses^{12,18,65}.

Improved modelling of archaeal gene family evolution using branch-heterogeneous phylogenetic reconciliation models

Early reconciliation analyses of the archaeal root made use of a limited set of archaeal diversity (e.g., 60 genomes in Williams et al.¹⁸), and a model that made the simplifying assumption that rates of duplication (D), transfer (T) and loss (L) do not vary across the branches of the species tree. However, various lines of evidence suggest that D, T and L rates do vary among archaeal clades, reflecting differences in the fixation of D, T and L events among lineages. Note that here and in what follows, DTL rates and events refer to fixations in the population (which are detectable from gene tree-species tree reconciliation), not rates at which Ds, Ts or Ls first appear in a population; the latter would require population-genetic data estimates. Thus, DTL rates have the same interpretation as substitution rates in traditional phylogenetic models. Our analyses indicate that duplications appear to be more common in Asgard archaea^{18,76}, while losses are more common among DPANN lineages¹⁸. In principle, a failure to accommodate these varying gene fixation rates might mislead phylogenetic analyses that rely on reconciliation methods, just as failure to accommodate varying evolutionary rates or compositions among alignment sites can induce high support for the incorrect trees in traditional sequence-based phylogenetics^{58,77,78}. Both of these limitations can now be overcome: the genome sampling of the archaeal domain is far richer than previously, and we have recently developed more flexible phylogenetic reconciliation models that relax the assumptions of constant D, T and L rates across the species tree implemented in AleRax⁵⁶.

To begin, we fit a three-parameter branch-homogeneous reconciliation model (one D, T and L rate, respectively, common to all branches, referred to as global in what follows) to the dataset of 257 archaeal genomes and 7059 gene family phylogenies, and used an AU⁶⁴ test to compare support for a range of potential root positions that were also assessed in the outgroup rooting analyses (Fig. 1, 1–15). The result provided strong support for a root between DPANN Cluster 2⁶¹—a clade within DPANN—and all other archaea (referred to as DPANN Cluster 2 root, Fig. 1A) (Supplementary Data 9). To gain insight into the nature of the signal for the DPANN Cluster 2 root, we investigated the properties of the gene families that favoured the DPANN Cluster 2 root over alternative root positions, using gene family filtering to explore root signals¹². We noticed that the preference of DPANN Cluster 2 root came from gene families with a small number of species and a biased taxonomic representation (Fig. S8A, B), consistent with a “small genome attraction (SGA)” artefact, where the model prefers a root that separates smaller genomes from larger ones on the tree¹⁸. In principle, such a bias might arise in analyses under a branch-homogeneous DTL model if gene absences in genome-reduced lineages are over-interpreted as evidence that a given lineage never had a gene,

overlooking the possibility that genes might be absent due to secondary loss. Consistent with this possibility, DPANN Cluster 2 clades predominantly comprise symbionts with reduced genomes, suggesting that they might exhibit a higher-than-average gene loss rate²⁶. To test whether the assumption of a single, global loss rate might drive support for the DPANN Cluster 2 root, we added one more parameter to allow for a different loss (L) rate for DPANN archaea (DPANN_L), and reassessed root supports using an AU test⁶⁴. Adding a DPANN-specific loss rate substantially improved reconciliation model fit (BIC: 3284636.6, Supplementary Data 10), and, as predicted, the DPANN loss rate was inferred to be higher than that for the other clades (DPANN: 0.40–0.45, other clades: 0.26–0.27). Interestingly, this analysis expanded the set of plausible roots that could not be rejected by the AU test under the global model, including two alternative root positions that are located nearby on the unrooted tree: Euryarchaeota root or TACK +Asgard root ($P > 0.174$; Supplementary Data 11), suggesting an amelioration of SGA in the analysis.

The best-fitting reconciliation model supports a root at or near the base of Euryarchaeota

Given the shift in root support when explicitly modelling the higher loss rate in DPANN, we next explored a series of more complex reconciliation models with additional lineage-specific D, T and L parameters. Our aim was to identify a model that adequately captured lineage-specific variation in D, T and L rates across Archaea by using the BIC to assess model fit and to prevent overfitting, while also evaluating how root support varies when different aspects of genome evolution are taken into account (Supplementary Note 2).

These analyses indicated that the best-fitting reconciliation model for these data was a model in which each of the major clades/phyla of Archaea had their own gene-family D, T and L rates (Supplementary Data 10, assessed by BIC). The final model had 72 parameters (DTL_br2), with lineage-specific DTL rates for various Euryarchaeota, TACK+Asgard and DPANN archaeal lineages (see Methods). Under this 72-parameter model (Supplementary Data 10), an AU test rejected all roots except the Euryarchaeota root and a root within Euryarchaeota (between a clade comprising Methanobacteriota_A(B), Hydrothermarchaeota and Hadarchaeota, hereafter MHH root), both with a monophyletic DPANN emerging as sister to the Asgard+TACK group (Fig. 1B, $P \geq 0.48$, Supplementary Data 12). Support for the Euryarchaeota root and the MHH root under the branch-heterogeneous model was broad and particularly concentrated in the most widespread and taxonomically balanced families (Fig. S8I), with the Euryarchaeota/MHH root remaining the maximum likelihood root until the top 2% most widespread and balanced families had been filtered out of the analyses (Fig. S8I), in contrast to the gene family-removal analysis described above which found support for the DPANN Cluster 2 root derived from gene families with limited taxa representation present mainly on one side of the root (Fig. S8B). As another experiment, we further increased model complexity on the 257-genome dataset by using additional sets of DTL parameters for groups within the MHH clade (an 81-parameter model—see Supplementary Note 2). This further improved model fit slightly as judged by BIC (BIC: 3264897.57) and placed the root on the branch leading to

Euryarchaeota (AU test, $P = 0.99$, Supplementary Data 13), providing evidence against the adjacent MHH root (AU test, $P = 0.024$, Supplementary Data 13). However, root support across gene families was less robust under this model, with the increased confidence for the Euryarchaeota over the MHH root deriving from patchily-distributed gene families (Supplementary Note 2), potentially resulting from the challenge of adequately modelling both relatively recent, lineage-specific evolution and DTL regimes on the deepest branches of the tree. Our conservative conclusion from these results, based on both improved model fit and the robustness of support across families in the 72-parameter model, is that the Euryarchaeota and MHH roots are the best supported roots from our reconciliation analyses (Supplementary Note 2).

Given the improved model fit and change in root support observed with the better-fitting, branch-heterogeneous reconciliation models on the 257-genome dataset, we also reanalysed the 60-genome dataset from Williams et al.¹⁸, which had provided support for a DPANN root when analysed with an earlier branch-homogeneous reconciliation model. As in our focal dataset, allowing D, T and L parameters to vary across clades substantially improved model fit (BIC: 462712.66, Supplementary Data 14) and shifted support from the original DPANN root to a root between Euryarchaeota and all other Archaea ($P = 1$, Supplementary Data 15, Supplementary Note 3). Taken together, our results on new and previous datasets demonstrate that improved modelling of variation in D, T and L rates across the tree substantially improves reconciliation model fit and provides support for an archaeal root at or near the base of Euryarchaeota (Fig. 1B).

Dynamics of archaeal genome evolution

Given the impact of model complexity on root inference, we evaluated the effect of model parameterisation on ancestral reconstructions. Ancestral reconstructions are highly similar across different models (Pearson's r : >0.98 ; Supplementary Note 4, Fig. S9). Therefore, we performed ancestral reconstructions on the 72-parameter model with the origination rate of deep nodes optimised (see "Methods"). Our analyses suggest that archaeal gene family evolution has been mostly vertical, with a mean verticality (i.e., the proportion of genes at the bottom of a branch whose ancestry traces straight up through the same branch) estimated to be 79% (Supplementary Data 16). Transferability of gene families varies significantly across functional categories under both the Euryarchaeota root and the MHH root ($H = 885.19$, $P = 1.13 \times 10^{-174}$ and $H = 878.0$, $P = 3.75 \times 10^{-173}$, degree of freedom: 20, Kruskal–Wallis test, Supplementary Data 17 and 18), with genes involved in defence mechanisms (such as restriction endonuclease, COG category V), and cell wall biogenesis (such as glycosyltransferase, COG category M), being the most frequently transferred, and those involved in cell cycle and informational processing the least (Fig. S10 and Supplementary Data 19), in line with previous findings^{18,79}. Gene transfer has a strong impact on archaeal evolution: it outnumbers duplication events across nearly all (507/501 out of 513 branches for the Euryarchaeota/MHH root), and loss events on many branches (177/154 out of 513 branches). Specifically, we found that gene transfers are on average 10.87–11.18 times more frequent than duplication events in archaeal evolution (Supplementary Data 20), in agreement with the finding that transfers are more frequent than duplications in prokaryotic genomes^{80,81}. In line with previous studies^{18,76}, Asgard archaea had the highest inferred duplication rate compared to the other archaeal clades (Fig. 2A, B), whose duplication events (normalised by speciations) appeared significantly higher than its sister lineage Thermoproteota (Conover's test, $P < 0.012$, Supplementary Data 21), while the transfer and loss events of Asgard archaea (normalised by speciations) did not appear to be significantly different from what we observe for other archaea with the exception of DPANN lineages (Fig. 2A, B, Conover's test, Supplementary Data 22 and 23). This suggests that the eukaryote-like propensity to fix duplicate genes

had already begun to develop in our Asgard ancestors prior to the origin of eukaryotes.

Our analyses inferred significantly higher normalised loss and transfer events in DPANN Cluster 1 and DPANN Cluster 2 archaea compared to euryarchaeal and thermoproteotal lineages (Conover's test, $P < 0.05$, Supplementary Data 22 and 23), suggesting that genome evolution of DPANN was shaped by gene exchanges and losses, consistent with their reduced genomes, predicted symbiotic lifestyle and host dependencies. In fact, the branch leading to the common ancestor of DPANN archaea (LDCA) experienced 639.82/592.64 loss events, followed by another 490.19/588.91 events on the branch leading to Undinarchaeota and DPANN Cluster 2 archaea under the Euryarchaeota/MHH root (Supplementary Data 20; Figs. S11 and S12). Notably, while the duplication is significantly lower in DPANN Cluster 2 archaea, SpSt-1190 lineages appeared to have the highest normalised duplication events among DPANN Cluster 2 archaea (Fig. 2A, B and Supplementary Data 21). Additionally, we observed a high number of duplication and acquisition events (transfers and originations) at the branch leading to Halobacteria, whose members are suggested to have evolved from a methanogenic ancestor into aerobic/light-harvesting heterotrophs⁸² and are reported to be polyploid⁸³ with duplicated genes predicted to be involved in the osmotic stress response for adaptation to hypersaline environments (Figs. S11 and S12).

The Last Archaeal Common Ancestor was a (hyper-)thermophilic methanogen

Branch-wise reconciliation analysis provides an inference of gene family evolution and an estimate of ancestral genome content across the tree, allowing the reconstruction of the metabolic potential of LACA. Compared to previous work¹⁸, this reconstruction sampled broader archaeal diversity and made use of a more realistic, better-fitting reconciliation model to reconstruct archaeal genome evolution according to the BIC (Supplementary Data 10). On the basis of the relationship between number of arCOG gene families and genome size in modern archaeal genomes (Asgard archaea: Pearson's correlation coefficient: 0.967, $P = 3.27 \times 10^{-12}$, all other Archaea: 0.9587, $P = 2.57 \times 10^{-130}$, Fig. S13 and Supplementary Data 24), we estimated the genome size of LACA to having been 1.58 Mb (95% bootstrap confidence interval, BCI: 1.54 Mb to 1.61 Mb) for the Euryarchaeota root and 2.06 Mb (95% BCI: 2.01 Mb to 2.11 Mb) for the MHH root (Fig. 2C, D). Based on the inferred presence probabilities (PP) of all arCOG families, we further predicted that LACA encoded between 800 to 1219 arCOG gene families for the Euryarchaeota root, and 1190 to 1709 gene families for the MHH root (Supplementary Data 25 and 26). This estimate is obtained by integrating over the LACA PPs for all gene families: some gene families have LACA PPs equal to zero, suggesting that they originated later in archaeal evolution (e.g., ammonia monooxygenase subunit A, arCOG08676; for succinctness, PP values of Euryarchaeota/MHH root are reported as x/y unless specified otherwise; here, 0/0); others have moderate root PPs (≥ 0.5), suggesting they may have been present early in archaeal evolution (e.g., glucokinase, arCOG04280 with PP = 0.54/0.54), while others have high root PPs (≥ 0.75), representing strong evidence that the gene was already present in LACA (e.g., DNA-directed RNA polymerase subunit A, arCOG04527, PP = 0.99/1). Due to phylogenetic uncertainty, stochastic error and the mosaic ancestry of genomes, PP support for the constituent genes of a metabolic pathway is often variable. For example, while most canonical genes encoding proteins involved in replication^{84,85} (Fig. S14; Supplementary Data 25 and 26) and translation machinery⁸⁶ (Fig. S15; Supplementary Data 25 and 26) were inferred to have been present in LACA, some had low support. This incompleteness of certain modern metabolic pathways is expected, given the evolutionary changes through time and limitations of our analysis pipeline, including phylogenetic noise. Hence, we interpret our reconstruction of LACA's

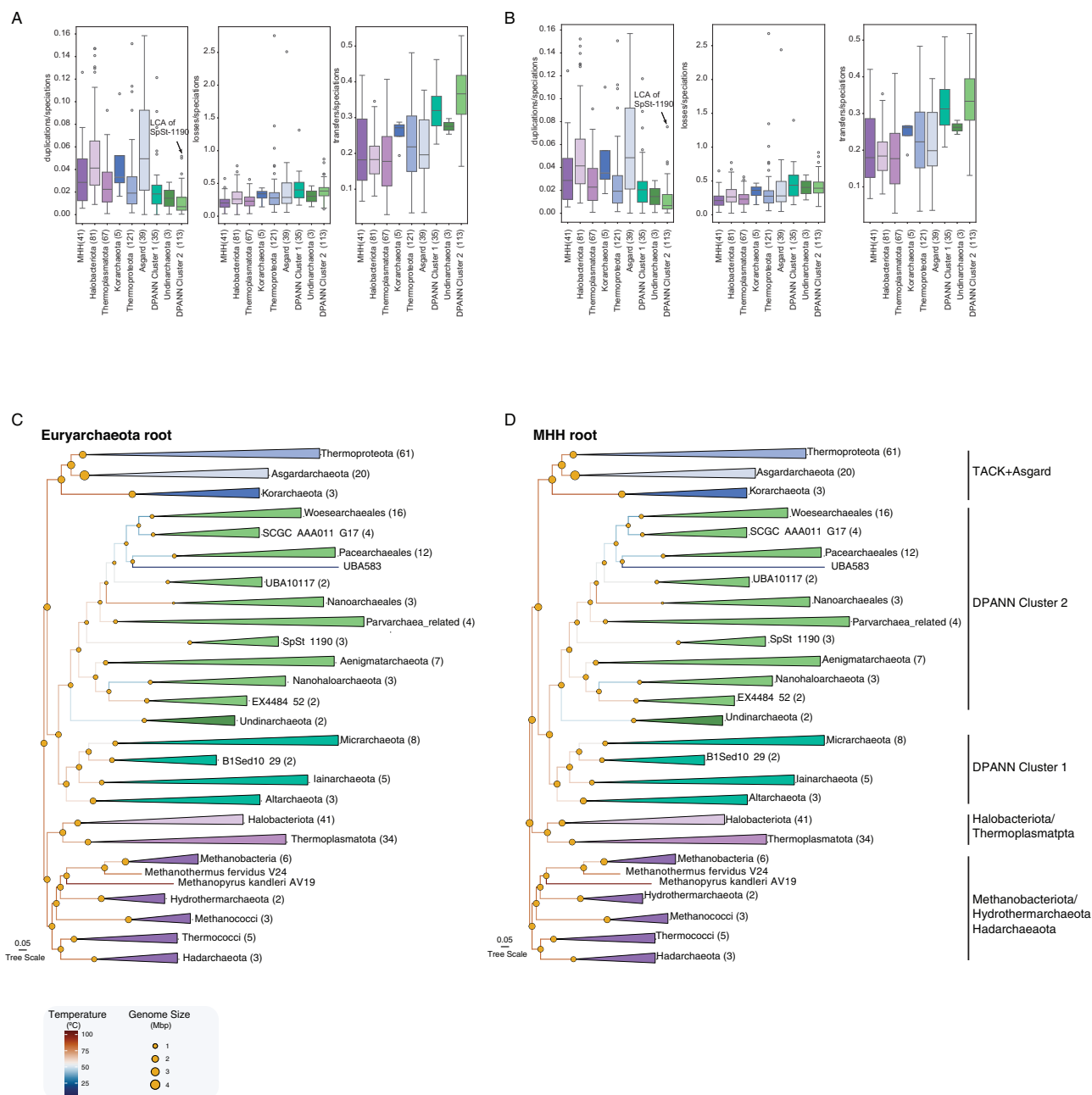


Fig. 2 | Evolution of gene family repertoires and temperature preferences in the Archaea. Boxplot of normalised duplication, loss and transfer events inferred at all branches of different archaeal groups (Methanobacteriota_A+Methanobacteriota_B+Hydrothermarchaeota+Hadarchaeota (MHH): $n = 41$, Halobacteriota: $n = 81$, Thermoplasmatota: $n = 67$, Korarchaeota: $n = 5$, Thermoproteota, $n = 121$, Asgardarchaeota: $n = 39$, DPANN Cluster 1: $n = 35$, Undinarchaeota: $n = 3$, DPANN Cluster 2: $n = 113$) under the **A** Euryarchaeota root and **B** MHH root. The central line indicates the median, and the box limits represent the interquartile range (IQR) (25th to 75th percentiles). Whiskers extend to the most extreme data points within 1.5× IQR from the box. Data points beyond the whiskers are outliers. A Kruskal–Wallis rank sum test (one-sided) was performed to assess differences among archaeal groups. Post hoc pairwise comparisons were conducted using the Conover's test.

Significant results are provided in Supplementary Data 21, 22 and 23. The number in brackets corresponds to the number of included nodes and tips per taxonomic group. Genome sizes appear to have remained approximately constant in Euryarchaeota (including Halobacteriota, Thermoplasmatota, Methanobacteriota_A, Methanobacteriota_B, Hydrothermarchaeota and Hadarchaeota) and TACK, expanded in Asgard archaea, and generally reduced in DPANN (including DPANN Cluster 1, Undinarchaeota and DPANN Cluster 2 archaea), with some lineage-specific exceptions. The last archaeal common ancestor might have been a (hyper-) thermophilic organism, encoding reverse gyrase under both the **C** Euryarchaeota and **D** MHH root hypotheses. The inferred temperature is provided in Supplementary Data 29.

metabolism based on the inference of pathways using the same PP cutoff (≥ 0.5) as in previous analyses^{12,18,62}.

The inferred gene content of LACA suggests that it already had most of the components of modern archaeal DNA replication, transcription, translation and protein quality control machinery, in agreement with previous work¹⁸ (Figs. S14–S18). We also inferred an

FtsZ-based cell division system, components of secretion systems and various transporter genes (Supplementary Data 25 and 26). Additionally, we identified most of the proteins required for archaeal biosynthesis⁸⁷ and an archaeal/vacuolar-type ATP synthase (Supplementary Data 25 and 26; Fig. S19). The improvement of the reconciliation model and expanded archaeal diversity furthermore allowed us

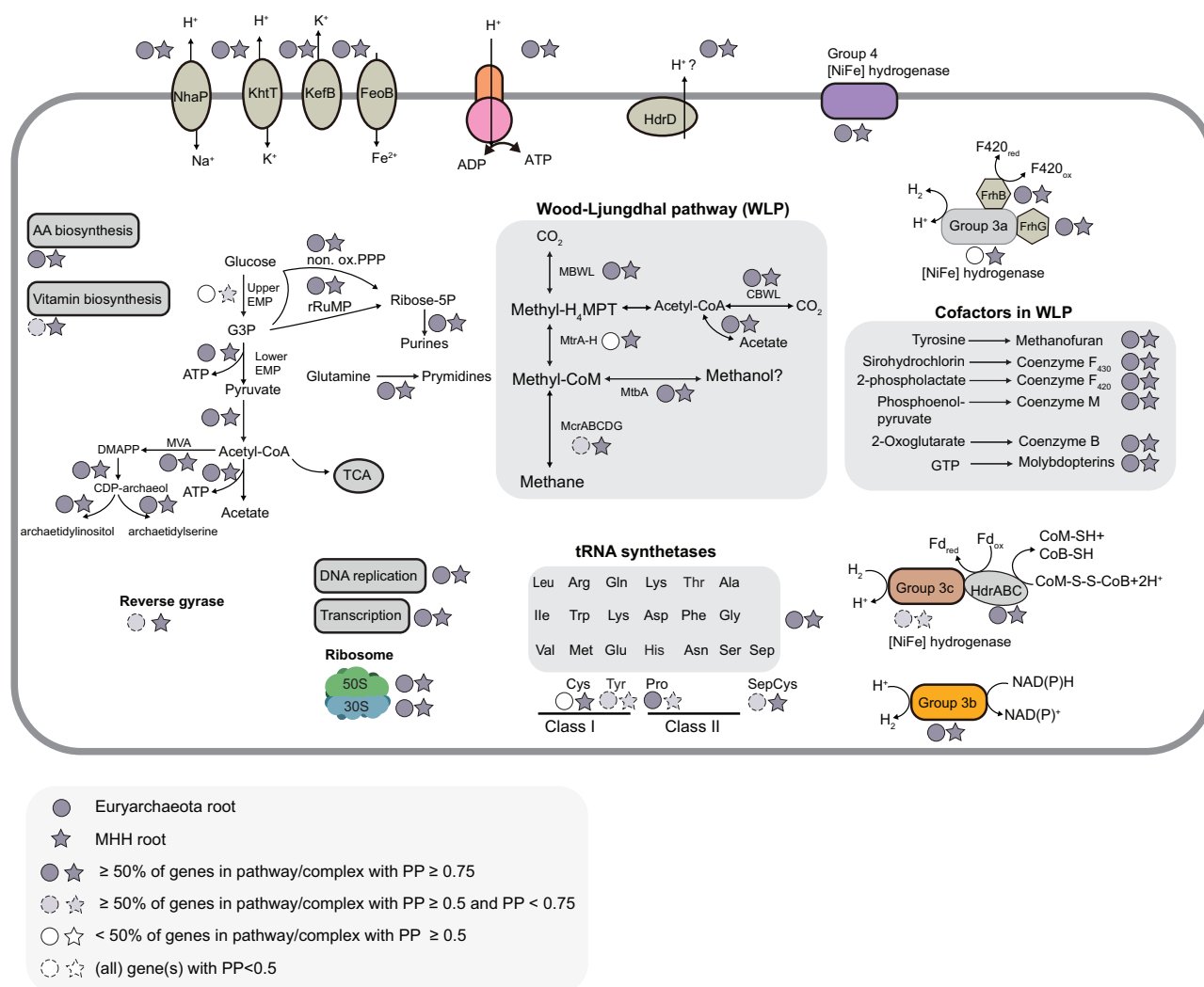


Fig. 3 | Ancestral reconstruction of the last archaeal common ancestor (LACA) under the two most plausible root hypotheses. The reconstruction is based on genes that are mapped to LACA with a presence posterior probability (PP) ≥ 0.5 under the two most plausible root positions. Pathway is shown by arrows, and complexes/genes by boxes. We recovered strong support for methyl-coenzyme M reductase in LACA, along with the carbonyl- and methyl-branch of the Wood-Ljungdahl pathway, suggesting LACA was a methanogen. Posterior probability for genes used in reconstruction is provided in Supplementary Data 25, 26, 27 and 28. A more detailed figure is provided in Figs. S19 and S20. The completeness of amino

acid and vitamin biosynthesis genes is counted from gene families that are inferred to have a presence probability ($PP \geq 0.5$) in at least two ancestors (Supplementary Data 25 and 26). MBWL methyl branch of Wood-Ljungdahl pathway, CBWL carbonyl branch of Wood-Ljungdahl pathway, TCA tricarboxylic acid cycle, G3P glycerol-3-phosphate, MVA mevalonate pathway, EMP Embden-Meyerhof-Parnas pathway, DMAPP Dimethylallyl pyrophosphate, Non. ox. PPP: non-oxidative pentose phosphate pathway. The image of the ribosome is adapted from the ribosome icon by Jaiganesh <https://github.com/jaiganeshjg> is licensed under CC0, <https://creativecommons.org/publicdomain/zero/1.0/>.

to map the presence of many additional components to LACA, including proteins involved in the biosynthesis of key cellular building blocks, such as nucleotides and lipids, which have been missed previously¹⁸. For example, a complete mevalonate pathway (MVA) via anhydromevalonate 5-phosphate, along with the genes for archaeal biosynthesis, including glycerol-1-phosphate dehydrogenase (GldA, arCOG00982, $PP = 1/0.99$) (Fig. 3; Fig. S19, Supplementary Data 25 and 26), were mapped back to LACA. The improvement also included many genes in de novo purine and pyrimidine and cofactor biosynthesis (Supplementary Note 5).

Further, we identified almost all components of the lower glycolytic, non-oxidative pentose phosphate and ribulose monophosphate pathways with high PP support (Fig. 3; Fig. S19, Supplementary Data 25 and 26). Several genes in the tricarboxylic acid (TCA) cycle were also mapped to LACA. (Fig. S19; Supplementary Data 25 and 26). Among the various carbon fixation pathways operating in archaea^{88–92}, the Wood-Ljungdahl pathway (WLP) has been suggested to represent

an ancient metabolic pathway^{18,93,94} potentially tracing back to the last universal common ancestor^{62,95,96}. In line with this, our results traced almost all components of the tetrahydromethanopterin methyl- and carbonyl-branch of the WLP to LACA (Fig. 3; Fig. S20, Supplementary Data 25 and 26). Our reconstructions also indicated the presence of nearly all subunits of the key enzyme of methanogenesis/methane oxidation, i.e., methyl(acyl)-coenzyme M reductase (M(A)cr) (McrA, arCOG04857, $PP = 0.5/0.58$, McrB, arCOG04860, $PP = 0.61/0.56$, McrG, arCOG04858, $PP = 0.39/0.82$, McrD, arCOG04859, $PP = 0.78/0.96$, and McrC, arCOG03225, $PP = 0.74/0.94$) (Fig. 3; Supplementary Data 25 and 26) in LACA. Moreover, we identified many methanogenesis marker proteins and genes involved in methanofuran, coenzyme F₄₃₀, coenzyme F₄₂₀, coenzyme B, coenzyme M and molybdopterins biosynthesis in LACA (Fig. 3; Fig. S20, Supplementary Data 25 and 26). This provides the first support for a methanogenic LACA based on ancestral genome reconstructions, in agreement with the geological record⁹⁷. LACA was furthermore inferred to have encoded two out of

eight subunits of the tetrahydromethanopterin S-methyltransferase (Mtr) complex (MtrA, arCOG03221, PP = 0.92 and MtrH, arCOG04336, PP = 0.98) under the Euryarchaeota root (Fig. 3; Fig. S20 and Supplementary Data 25), and six out of eight subunits under the MHH root (MtrA, arCOG03221, PP = 0.99; MtrB, arCOG04867, PP = 0.65; MtrC, arCOG04868, PP = 0.58; MtrE, arCOG04870, PP = 0.6; MtrG, arCOG03380, PP = 0.71; MtrH, arCOG04336, PP = 0.97) (Fig. 3; Fig. S20, Supplementary Data 26). This enzyme is used by many hydrogenotrophic methanogens for sodium ion translocation^{40,98}. High support for the presence of methanol-corrinoid protein: coenzyme M methyltransferase (MtbA, arCOG03325, PP = 0.96/0.94 and methanogenic corrinoid protein (MtbC, arCOG02028, PP = 0.77/0.99) at the root suggests that LACA might have been able to use methylated compounds for methanogenesis. In addition, acetate-CoA ligase (AcdA, arCOG01340, PP = 0.60/0.85) and acetyl-CoA synthetase (Acs, arCOG01529, PP = 0.94/0.97) were also recovered at the root, indicating the potential to metabolise acetate. Thus, our analyses indicate the antiquity of the WLP and methanogenesis in Archaea^{39,99–101}, with the potential for CO₂-reducing methanogenesis being root-dependent.

Our reconstruction indicated high PP support for a group 3 [NiFe]-hydrogenase (PP = 0.98/0.99) (Supplementary Data 27 and 28, Figs. S20, S21, Supplementary Note 5), among which we were able to map a group 3c [NiFe]-hydrogenase (PP = 0.66/0.68) (Supplementary Data 27 and 28) and heterodisulphide subunits (HdrA, arCOG02236, PP = 0.73/0.81; HdrC, arCOG00964, PP = 0.87/0.99) to LACA with at least moderate support (Supplementary Data 25 and 26). This suggests that LACA might have had the ability to use a group 3c [NiFe]-hydrogenase to bifurcate electrons from H₂ to reduce ferredoxin and heterodisulphide of coenzyme B and M, coupled with the Mcr complex to form methane. Furthermore, our analyses indicated that an ancestral group 4 [NiFe] hydrogenase might have been present in LACA (PP = 0.9/0.99; Supplementary Data 27 and 28; Fig. S20). Group 4 [NiFe]-hydrogenases comprise various respiratory membrane-bound enzymes in modern organisms and may have been involved in energy conservation in the last universal common ancestor^{102–104}. These results are consistent with the view that [NiFe]-hydrogenases are ancient enzymes¹⁰³. Additionally, we recovered a subunit of heterodisulphide reductase (HdrD, arCOG00333, PP = 0.99/0.97), which is present in various modern archaea, including methanogens^{105,106}, and might facilitate electron transfer.

Finally, we estimated the optimal growth temperature of LACA based on the observation that the amino acid composition of our 45-gene supermatrix showed a significant correlation with the optimal growth temperature (Pearson's r : -0.87, $P = 8.951 \times 10^{-20}$) of modern Archaea (Fig. S22). We used the branch-heterogeneous Coala substitution model¹⁰⁷ to sample 100 ancestral sequences reconstructed at nodes of the trees and inferred the temperature optima for each node⁴⁸. The analysis suggested that LACA may have been a (hyper-) thermophile (median 78.78°C, 95% CI: 74.15–83.52°C under the Euryarchaeota root; median 79.40°C, 95% CI: 74.18–84.30°C under the MHH root; Supplementary Data 29), consistent with previous work^{18,107} and with the inference that reverse gyrase, a specific hallmark of (hyper-) thermophiles, may have been present in LACA under the Euryarchaeota/MHH roots (arCOG01526, PP = 0.68/PP = 0.75, Supplementary Data 25 and 26).

The DPANN ancestor was a free-living organism

Our rooted species tree resolves DPANN as a monophyletic sister clade of TACK and Asgard. We reconstructed a broadly similar metabolic repertoire of the LDCA and the last TACK+Asgard common ancestor, with the LDCA appearing to encode fewer genes in vitamin biosynthesis (Supplementary Note 6; Figs. S19 and S20). Further, the LDCA under the Euryarchaeota and MHH root, respectively, was estimated to have encoded a genome of 1.32 Mb/1.68 Mb (95% CI: 1.28 Mb to 1.36 Mb / 95% CI: 1.64 Mb to 1.72 Mb), with 546–827/820–1152 arCOG families

(Supplementary Data 24). This is comparable to modern Altiarchaeota genome sizes and gene repertoires (1.17 to 2.41 Mp; 1219 to 2776 total coding sequences) (Supplementary Data 1), suggesting that the LDCA may have been a free-living organism similar to modern Altiarchaeota^{27,108}. Indeed, our inference of the ancestral metabolic content of the LDCA is compatible with a free-living lifestyle. For instance, our analyses indicate the presence of almost all components of archaeal ether lipid biosynthesis and de novo purine and pyrimidine biosynthesis pathways in the LDCA (Fig. 4; Supplementary Data 25 and 26). A number of genes involved in vitamin and amino acid biosynthesis were also recovered (Fig. 4; Supplementary Data 25 and 26). This contrasts with many DPANN Cluster 2 representatives, including co-cultivated symbionts such as members of Nanoarchaeales^{34–36,109} like *Nanoarchaeum equitans*^{29,110}, and members of Nanohaloarchaeota^{30,31}, such as *Nanohaloarchaeum antarcticus*³⁰, that do not encode lipid, nucleotide, amino acid or cofactor biosynthesis pathways. We also identified components of central carbon metabolic pathways, including the lower glycolytic pathway, pentose phosphate pathway (PPP) and TCA cycle in the LDCA (Fig. S19). Notably, although the M(A)cr complex was lost in LDCA and the last DPANN Cluster 1 common ancestor (LDC1CA), both ancestors were predicted to have encoded many genes in the H₄MPT methyl and carbonyl branch of WLP, as well as an acetate-CoA synthetase (Fig. S20). However, genes of the WLP pathway were lost on the branch leading to the last common ancestor of Undinarchaeota+DPANN Cluster 2 archaea and are absent in most modern DPANN archaea except Altiarchaeota and SpSt-1190^{27,111}.

We inferred an ongoing trend of convergent genome reduction during the evolution of DPANN archaea (Fig. 2C, D; Fig. S23 and Supplementary Data 24). In particular, DPANN Cluster 2 archaea are characterised by genome streamlining that appears to have occurred in parallel in different lineages^{61,112}. Specifically, our reconstructions inferred a reduction in genome size from 1.32/1.68 Mb in LDCA to 0.92/1.26 Mb in the DPANN Cluster 2 common ancestor (LDC2CA, Supplementary Data 24). The reduction in genome content included genes encoding enzymes of the WLP, ribulose monophosphate pathway (RuMP), vitamins, amino acids, nucleotides and lipid biosynthesis pathways, which indicates dependency on external sources by the LDC2CA (Fig. 4; Fig. S19, Supplementary Data 25 and 26, Supplementary Note 7). However, we identified most genes of the lower glycolytic pathway and acetate-CoA ligase in the LDC2CA, in line with experimental support indicating that this machinery may be indispensable for these symbionts in the absence of host-derived ATP¹¹³. We also inferred that most of the archaeal subunits⁸⁷ of the motor complex (ArlHII) were present in the LDC2CA, which were found in extant representatives of DPANN Cluster 2 archaea (Supplementary Data 25–26)^{113,114}. Similarly, we identified most of the components of DNA replication, transcription and translation machinery in LDC2CA and the last common ancestor of Nanoarchaeota, in line with extant representatives^{24,25,115,116}. In contrast to this general trend, we inferred that some clades within Woesearchaeales and Pacearchaeales experienced genome expansion (Fig. S23). For example, we inferred an increase of 0.34 to 0.39 Mb in one Woesearchaeales clade, and an increase of 0.26 to 0.27 Mb in one Pacearchaeales clade, compared to the last common ancestor of Woesearchaeales and Pacearchaeales, respectively (Fig. S23 and Supplementary Data 24). The genome content expansion of the Woesearchaeales clade includes genes involved in amino acid transport and metabolism, as well as energy production and conversion, in line with a previous inference¹¹². For example, several subunits (1–4) of the archaeal/vacuolar-type ATP synthase were inferred to have been re-acquired during this expansion (Supplementary Data 30). In contrast, the majority of genes that contributed to genome expansion in Pacearchaeales were inferred to be involved in translation, transcription, and replication machinery

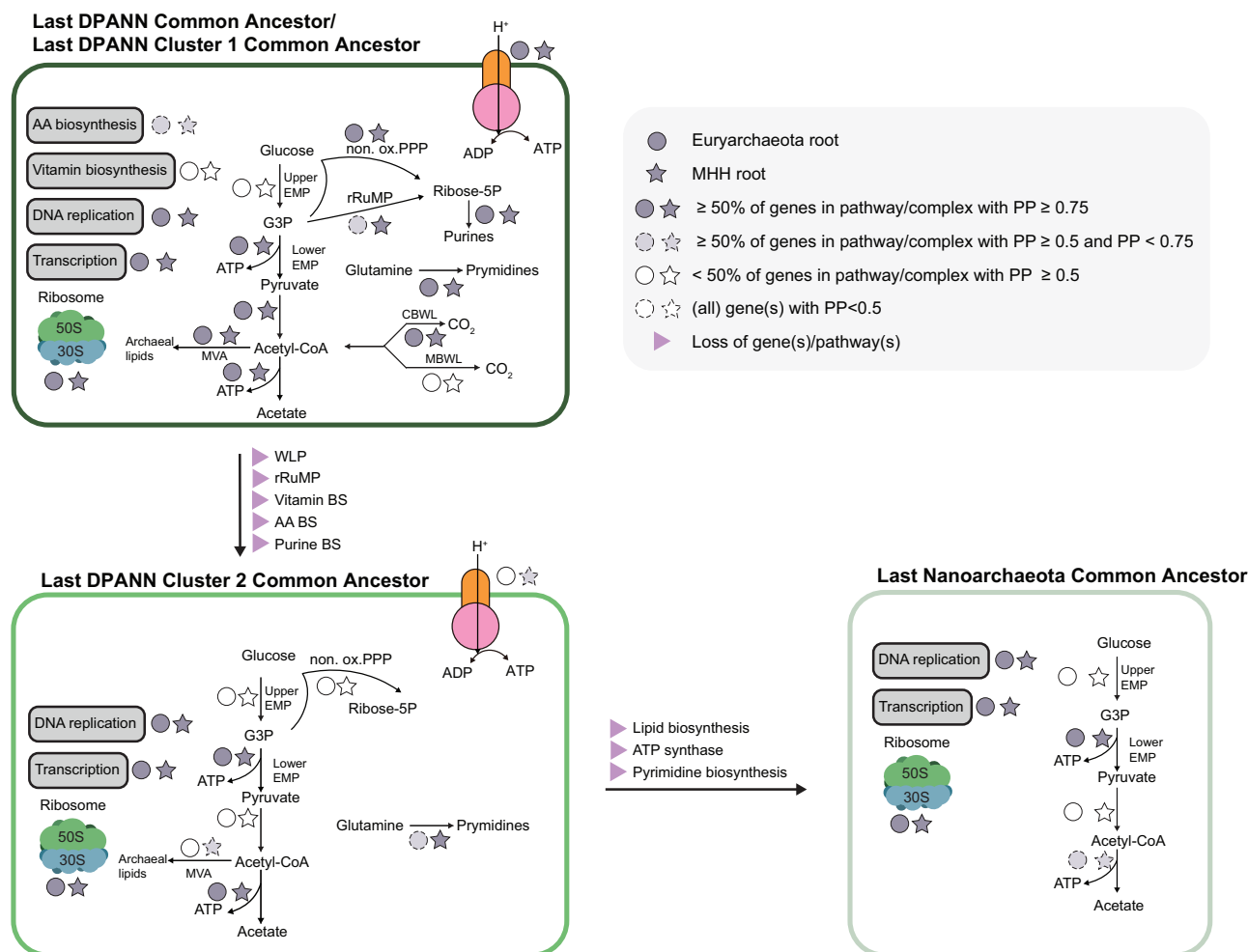


Fig. 4 | Ancestral reconstruction of the last common ancestor of DPANN archaea, DPANN Cluster 1 archaea, DPANN Cluster 2 archaea and Nanoarchaeota. The reconstruction is based on genes that could be mapped to these ancestors with a presence posterior probability (PP) ≥ 0.5 . The pathway is shown by arrows, and complexes/genes by boxes. We recovered strong support for the carbonyl- and methyl-branches of the Wood-Ljungdahl pathway (WLP), in addition to lipid-, vitamin- and amino acid-biosynthesis genes in the last common ancestor of DPANN and the last common ancestor of DPANN Cluster 1, suggesting a free-living lifestyle. However, genes related to the WLP, ribulose monophosphate pathway, vitamin and amino acid biosynthesis pathways were lost in the DPANN

Cluster 2 archaea, followed by a loss of genes comprising the ATP synthase, and lipid and nucleotide biosynthesis genes in the last common ancestor of Nanoarchaeota (Nanobdellota), increasing their host-dependency (Supplementary Note 7). Posterior probability for genes used in reconstruction is provided in Supplementary Data 25, 26, 27, and 28. G3P glycerol-3-phosphate, MVA mevalonate pathway, EMP Embden-Meyerhof-Parnas pathway, Non. ox. PPP: non-oxidative pentose phosphate pathway. The image of the ribosome is adapted from the ribosome icon by Jaiganesh <https://github.com/jaiganeshjg> is licensed under CC0. <https://creativecommons.org/publicdomain/zero/1.0/>.

(Supplementary Data 31). For example, acquisitions of transcriptional regulators were inferred under both the Euryarchaeota and MHH root (Supplementary Data 31) for the latter lineage.

Discussion

The archaeal phylogeny continues to be revisited and revised as genomes from new groups become available, and phylogenomic methods improve^{1,3,5,6,7,17,118}. While many studies consistently support the clanhood of DPANN and TACK+Asgard archaea^{13,14,18,19,24,25,39,62}, the clanhood of Euryarchaeota remains debated. Some analyses recover Euryarchaeota as a monophyletic group^{14,18,24,25,39,62,117}, whereas others suggest it to be paraphyletic, with (hyper-)thermophilic lineages, Thermococci and/or Hadarchaeota, branching basal to TACK+Asgard^{17,19,48,119}. These two lineages were proven difficult to place in previous work^{13,14,17,119–121}, and are placed at the base of Euryarchaeota in the phylogeny summarised from all gene trees (Fig. S4). Nevertheless, our analyses consistently recovered the clanhood of Euryarchaeota under the progressive removal of compositionally biased sites, under the best-fitting MAST model⁶⁹, and from the summary of all gene trees

using Astral-Pro 3⁶⁸, in line with a previous observation in which support for the clanhood of Euryarchaeota emerged upon site filtering¹⁷. While our analysis recovered Hydrothermarchaeota within Methanococci, Methanopyri and Methanobacteria with low support UFBOOT/SH-aLRT support (21.6/83) in the untreated dataset (Fig. 1), an AU test could not reject a topology in which Methanococci, Methanopyri and Methanobacteria are sister to Hydrothermarchaeota ($P=0.5$). This is consistent with previous work either recovering the latter topology^{17,99,119,122} or placing Hydrothermarchaeota within Methanobacteriota^{13,61}. While some studies placed Altiarchaeota at the base of all DPANN archaea^{17,19,48}, our analyses supported the placement of Altiarchaeota within DPANN as a member of DPANN Cluster 1 once gene families strongly affected by horizontal evolution (including interdomain transfer) were removed, suggesting that the alternative placement was induced by including non-vertical markers (Supplementary Note 1). Finally, although many previous studies placed the Korarchaeota inside or basal to Thermoproteota^{19,21,123–125}, we recovered Korarchaeota as a sister lineage of a monophyletic clade of TACK+Asgard after the removal of non-vertical markers (Fig. S6). This

placement was consistent across analyses, including the progressive removal of compositionally biased sites, the application of the best-fitting MAST model, and gene tree summary approach using Astral-Pro 3, in agreement with recent work^{61,78,126,127} and the latest assignment of Korarchaeota as a phylum-level lineage in GTDB¹²⁸.

We applied various methods to root the archaeal tree and assessed the confidence of various root placements⁵¹. Our results reveal that outgroup rooting and non-reversible models are unable to distinguish among a broad set of potential root placements (Supplementary Data 3 and 6). By contrast, the GFMix model, which accounts for variation in amino acid frequencies across sites and branches, supported a Halobacteriota+Thermoplasmata root, in line with a previous study using a divide-and-conquer phylogenomic approach^{19,20}. While these analyses are typically based on information from a small set of marker genes, phylogenetic reconciliation allows to incorporate additional information based on gene evolutionary histories^{12,18,57,65,129}. Earlier attempts to resolve the root of Archaea using gene tree-species tree reconciliation approaches were based on a small taxon set and relied upon a simple reconciliation model that did not incorporate variation in rates of duplication, transfer and loss between different lineages¹⁸—a limitation that can induce artefacts, especially when including a diverse set of reduced genomes from symbionts. Analyses using better-fitting models placed the archaeal root on or within Euryarchaeota, with DPANN emerging as a sister group to TACK+Asgard. The root recovered in the GFMix analysis was one node away from these two adjacent branches, defining a small “root region” on the archaeal tree at or near the base of the Euryarchaeota. These results emphasise the importance of integrating parameter-rich, realistic models of evolution that improve model fit and resolve deep archaeal phylogeny.

By including only sequences from archaea and avoiding the need for a bacterial outgroup, the reconciliation approach enables many more gene families to be used to resolve the archaeal root, and reduces the impact of long branch attraction artefacts on the results. However, a limitation of the archaeal focus is that inter-domain gene transfers between archaea and bacteria in individual arCOG gene families will not be detected, as the reconciliation method only sees the archaeal subfamily. Our analyses indicated that the same root is recovered on that subset of gene families least prone to inter-domain transfer (Supplementary Data 32), suggesting that this does not cause issues in practice. In principle, a reconciliation analysis including both prokaryotic domains would provide a higher-resolution map of gene family origins (though still limited by genome sampling of extant lineages and the now-extinct lineages in which gene families most likely originated), albeit at the expense of gene family tree and root inference accuracy.

Our reconciliation approach simultaneously allowed reconstructing ancestral gene sets along the rooted archaeal tree, resolving a relatively complex LACA that was likely a (hyper-)thermophilic methanogen^{18,84–86,130,131}. The inferred metabolic content of LACA is consistent with methyltrophic methanogenesis or potentially CO₂-reducing methanogenesis^{39,94,99,132}. We hypothesise that such a hyper-thermophilic ancestor could have thrived in environments analogous to hydrothermal vents or geothermal subsurface systems with strong reducing conditions, abundant in H₂, CO₂ and methylated compounds such as methanol^{133,134}.

While various studies have recovered a deep split between DPANN and all other archaeal lineages^{14,18,24}, more recent work has supported DPANN forming a derived lineage within Euryarchaeota in a tree rooted between TACK+Asgard and all other archaea¹⁷. Our analyses also recovered DPANN as a derived lineage, albeit branching sister to the TACK+Asgard (Fig. 1). These results suggest that the symbiotic lifestyle of DPANN archaea is a secondary adaptation rather than an ancestral trait. In fact, our analysis suggests that the LDCA might have been a free-living organism, with an ancestrally reduced vitamin biosynthesis

repertoire, potentially representing an early step towards dependency within a microbial community, and that subsequent diversification of DPANN lineages was shaped by high rates of gene loss and transfer. While genome reduction in both DPANN and CPR can thus be considered as a derived trait, their evolutionary trajectories differ. Reduction in CPR likely occurred early on the stem lineage¹², whereas DPANN gene loss was more episodic and, in part, occurred convergently in different lineages. However, genome reduction in the stem leading to DPANN Cluster 2 exhibits analogous patterns as observed for CPR, involving the loss of genes coding for proteins involved in central carbon and energy metabolism, including the ATP synthase, as well as amino acid biosynthesis¹². The loss of metabolic pathways is also a common feature in reductive genome evolution of bacterial endosymbionts of insects^{24,25,135}. In contrast to those endosymbionts, DPANN archaea appear to retain genes for proteins that are involved in DNA repair and transcription. This difference may reflect their distinct symbiotic strategies: all DPANN symbionts that have been successfully co-cultivated have episymbiotic stages^{29–31,33,113,136}. Therefore, they likely experience stronger selection to maintain key components for an information processing machinery than strictly vertically inherited intracellular endosymbionts. Furthermore, the ectosymbiotic lifestyle of DPANN archaea might explain the observed genome expansion in certain DPANN lineages, such as Woesearchaeales and Pacearchaeales, through high rates of horizontal gene transfer. Prospective analysis of the sources and mechanisms of these gene transfers may allow for identifying host organisms for the vast diversity of uncultivated DPANN.

Methods

Taxa selection

Contigs and protein files of archaeal genomes were downloaded from NCBI in August 2021 (n = 7,967). If protein files were unavailable, gene calling was performed using Prokka (v1.14, settings: --kingdom Archaea --addgenes --increment 10 --compliant --centre UU --norma --notrna)¹³⁷. Taxonomic classifications were assigned to all genomes using GTDB-Tk using GTDB release 202¹³⁸, and the original NCBI tax ID was converted into an NCBI taxonomy string using ncbi2tax2lin (<https://github.com/zyxue/ncbi2tax2lin>). Genome completeness was estimated using CheckM2 (v0.1.3)¹³⁹.

Genomes with completeness >50% and contamination <5% were used for an initial phylogenetic analysis using the PhyloSift marker set¹⁴⁰. In brief, homologues of 34 PhyloSift marker proteins were identified in each genome using the PhyloSift v1.0.1 “find” mode (settings: --besthit). Sequences were aligned with MAFFT (v7.407, settings: --reorder¹⁴¹, trimmed with BMGE (v1.12, settings: -t AA -m BLOSUM30 -b 2 -h 0.55)¹⁴². The trimmed alignments were concatenated using catfasta2phyml.pl (<https://github.com/nylander/catfasta2phyml>) and the concatenated protein alignment was used to reconstruct a phylogenetic tree using IQ-TREE (v1.6.7, settings: -m LG + G -nt AUTO -bb 1000 -alrt 1000)¹⁴³. Based on this tree, a sub-selection of 1,711 archaeal reference genomes was conducted based on high genome completeness, low contamination and taxonomic diversity.

To further thin out the taxon selection, two rounds of species tree inferences were performed using 51 universal, single-copy marker proteins suitable for archaeal phylogenetic analyses⁶¹. Specifically, a customised TIGR database was queried against a database generated from all archaeal proteins using hmmsearch (v3.1b2)^{144,145}. The output was parsed to only retain hits with an e-value ≤ 1E-3 and then to only include the best hit per protein selected using the lowest e-value and the highest bit score. Selected sequences were aligned with MAFFT (v7.407, settings: --reorder¹⁴¹, and alignments were trimmed with BMGE (v1.12, settings: -t AA -m BLOSUM30 -b 2 -h 0.55)¹⁴². Individual alignments were concatenated using catfasta2phyml.pl, and a phylogenetic tree was generated using IQ-TREE (v1.6.7, settings: -m LG + C20 + F + R -bb 1000 -alrt 1000)¹⁴³ and visualised using FigTree (v1.4.4).

The tree was manually inspected, and an additional taxa subset was selected based on genome completeness and contamination statistics while maintaining a taxonomically diverse set of genomes with a specific focus on the DPANN archaea. This workflow was repeated, resulting in a selection of 513 taxa that were further reduced to the focal taxa set of 257 genomes with balanced taxa representation of Euryarchaeota, TACK+Asgard and DPANN (Supplementary Data 1). Njordarchaeia was not included in the analyses owing to the timing of genome availability¹⁴⁶ (August 2021) and the difficulty of accommodating its compositional bias under conventional C-series models^{76,78}.

Genome annotations

For further functional annotation, proteins from 257 archaeal genomes were compared against several databases, including COGs (using the 2020 update)¹⁴⁷, arCOGs (version from 2020)¹⁴⁸, the KO profiles from the KEGG Automatic Annotation Server (KAAS; downloaded April 2019)¹⁴⁹, the Pfam database (Release 31.0)¹⁵⁰ and the TIGRFAM database (Release 15.0)¹⁴⁴. Individual database searches were conducted as follows by querying these databases against all archaeal genomes using *hmmsearch* v3.1b2 (settings: *-tblout -domtblout -notextw*)¹⁴⁵. The output was parsed to only include the best hit for each protein using the lowest e-value and the highest bit score value. Additionally, all proteins were scanned for protein domains using InterProScan (v5.29-68.0; settings: *-iprlookup -goterms*)¹⁵¹.

Inferring an unrooted archaeal species tree

The focal taxon selection was used to infer a maximum likelihood phylogeny using 51 universal, single-copy marker proteins as described above, but with minor adjustments to improve the quality of the species tree⁶¹—i.e., markers with more than 10% duplicated proteins were discarded, leaving 45 markers. Sequences were aligned using MAFFT or MAFFT-L-INS-i for gene families with more than or fewer than 1000 sequences, respectively, and trimmed with BMGE. Single gene trees were generated with IQ-TREE 2 (v2.1.2, settings: *-m LG -T AUTO -keep-ident -threads-max 2 -B 1000 -bnni*)⁶³. A custom script (sequence_correlator.py, <https://zenodo.org/records/18701544>) was run to identify multi-copy proteins¹⁵. Each single gene tree was visually inspected to identify the correct protein sequences per marker and remove long-branching sequences. After removing problematic sequences, trimmed alignments were generated using the procedure described above and concatenated using *catfasta2phym.pl* (<https://github.com/nylander/catfasta2phym.pl>). The best fitting phylogenetic model was estimated using IQ-TREE 2 (v2.1.2, settings: *-m MF -mset LG -madd LG + C10, LG + C10 + G, LG + C10 + R, LG + C10 + F, LG + C10 + R + F, LG + C10 + G + F, LG + C20, LG + C20 + G, LG + C20 + F, LG + C20 + G + F, LG + C20 + R, LG + C20 + R + F, LG + C30, LG + C30 + G, LG + C30 + R, LG + C30 + F, LG + C30 + R + F, LG + C30 + G + F, LG + C40, LG + C40 + G, LG + C40 + R, LG + C40 + F, LG + C40 + R + F, LG + C40 + G + F, LG + C50, LG + C50 + G, LG + C50 + R, LG + C50 + F, LG + C50 + R + F, LG + C50 + G + F, LG + C60, LG + C60 + G, LG + C60 + R, LG + C60 + F, LG + C60 + R + F, LG + C60 + G + F --score-diff all*), via the Bayesian Information Criterion¹⁵², and a phylogenetic tree was inferred using IQ-TREE 2 (v2.1.2, settings: *-m LG + C60 + F + R -B 1000 -alrt 1000*)⁶³ and visualised using *ggtree* (v3.14.0)¹⁵³. The species tree was re-rooted at 15 different key positions (positions from literature and between each major lineage) using a custom script (*root_w_ete3.py*, <https://zenodo.org/records/18701544>).

To assess the clade support for major archaeal lineages, compositionally biased sites were filtered from the alignment in a progressive manner (from 10–80%, in 10% increments) using *alignment_pruner.pl* (https://github.com/novigat/davinciCode/blob/master/perl/alignment_pruner.pl; settings: *--chi2_prune*). For each pruned alignment, a phylogenetic tree was inferred with IQ-TREE 2 (v2.1.2, settings: *-m LG + C60 + F + R -bb 1000 -alrt 1000*)⁶³. All tree topologies are highly congruent (Fig. S2; all trees share quartet similarity above

0.99¹⁵⁴), so we used the phylogenetic tree based on the full alignment for all further analyses.

To assess the impact of taxonomic balance at the phylum level on the topology of our inferred species tree, we downsampled the 257 archaeal dataset to 53 taxa, retaining 3 representatives for each GTDB phylum. The marker sequences of these taxa were realigned using MAFFT-L-INS-i¹⁵⁵ and trimmed using BMGE v1.2¹⁴² (*-m BLOSUM30 -h 0.55 -t AA*). The species tree was inferred under the best-fitting LG + C60 + F + R model using IQ-TREE v2.1.2⁶³. We also inferred a species tree using ASTRAL-Pro 3⁶⁸ (settings: *-r 16 -s 16*) from single-gene trees of 241 (present in more than 80% of the archaeal species) and 7059 (all families) arCOG gene families (See *arCOG gene families for ancestral gene content reconstruction*).

Further, we assessed the statistical support for the placement of Thermococci and Hadarchaeota, and the monophyly of Euryarchaeota, and Korarchaeota using the AU test⁶⁴ as implemented in CONSEL v0.2¹⁵⁶, and the best-fitting MAST model⁶⁹ for the 257 archaeal dataset. Specifically, we constructed 3 constrained topologies: Euryarchaeota paraphyletic with Thermococci branching sister to Hadarchaeota+TACK+Asgard (compatible with a recent study¹⁷), Hadarchaeota and Thermococci branched basal to other Euryarchaeota (compatible with the Astral-Pro 3 topology), Koarchaeota branches inside Thermoproteota (compatible with pre-GTDB 10 classification). Subsequently, we performed a constrained topological search using IQ-TREE v2.1.2 (*-g TOPOLOGY -wslr -m LG + C60 + F + R*). Six MAST models (TMIX{LG + FO + R, LG + FO + R, LG + FO + R, LG + FO + R} + T, TMIX{LG + FO, LG + FO, LG + FO, LG + FO} + R + T, TMIX{LG + F + R, LG + F + R, LG + F + R, LG + F + R} + T, TMIX{LG + F, LG + F, LG + F, LG + F} + R + T, LG + FO + TMIX{R, R, R, R} + T, LG + FO + R + T) were fit to the data using the unconstrained tree (Fig. S5), and the trees containing three alternative hypotheses via the *-te* option using IQ-TREE v3.0.1¹⁵⁷. The BIC identified TMIX{LG + FO + R, LG + FO + R, LG + FO + R, LG + FO + R} + T as the best-fitting model.

Reciprocal outgroup rooting of an archaeal species tree with bacteria

To perform outgroup rooting of archaea, we selected 25 vertically evolving marker proteins shared between bacteria and archaea¹³ (Supplementary Data 2). COG marker sequences were identified using *hmmsearch* v3.1.2 from 257 archaeal and 103 bacterial genomes. Alignments were generated using MAFFT-L-INS-i and trimmed using BMGE (*-m BLOSUM30 -h 0.55 -t AA*). Single gene trees were inferred based on LG + G using IQ-TREE 2 (v2.1.2) and inspected to remove long-branching sequences and false paralogues. Sequences were realigned using MAFFT-L-INS-i, trimmed using BMGE (*-m BLOSUM30 -h 0.55 -t AA*), and concatenated into a supermatrix. Finally, we inferred an ML tree under the LG + C60 + F + R model, which recovered a DPANN root (Fig. S1). Subsequently, we generated 14 additional root hypotheses by relocating the bacterial outgroup attachment point to the archaeal ingroup (Supplementary Data 3), and conducted maximum likelihood based tree inferences under these topological constraints using the LG + C60 + F + R model in IQ-TREE 2 (v2.1.2) (*-g root -wslr*). AU tests were performed using CONSEL v0.2¹⁵⁶ using 1,000,000 bootstrap samples.

Rooting of archaeal species tree using a non-reversible model

Additionally, we inferred maximum-likelihood phylogenies from the same concatenated alignment using three non-reversible models implemented in IQ-TREE 2 (v2.3.6, setting: *-m NONREV + R10/ NONREV + G/NQ.pfam+R10*). NONREV + R10 provided the best model fit to the data. The likelihood of trees being rooted on every branch of the ML tree from three models was evaluated using the “root-test” option, and AU tests were performed for those trees using the option: *-zb 10000 -au*. ML phylogenies inferred using these three models are shown in Fig. S7. The results representing the 15 root positions that

were tested with outgroup-rooting and reconciliation analysis are provided in Supplementary Data 6.

Rooting of archaeal species tree using a non-stationary model

We also used the GFMix model (v1.2)^{17,58,59} with LG + C60 + F + G model to assess the 15 different rooting hypotheses based on the same concatenated alignment. Site-wise likelihoods were computed using the: -l options, and AU tests were conducted using CONSEL v0.2¹⁵⁶ with 1,000,000 bootstrap samples. The GFMix rooting analyses were also performed on the alignment and treefile from a recent study¹⁷ by allowing the ratio of FIMNKY-to-GARP to vary across branches (settings: -d -l). Afterwards, an AU test was performed.

Gene tree/species tree reconciliations to the root archaeal species tree and reconstruct ancestral gene content

arCOG gene families for ancestral gene content reconstruction. Gene families for ancestral reconciliations were generated based on the arCOGs, a COG-based orthology database that captures archaeal protein family evolution¹⁴⁸. COG-based orthology has been shown to perform better than de novo clustering approaches in reconciliation analyses of bacterial genome evolution¹². To this end, proteins from an initial taxon set with 513 archaeal genomes (relatively large representation of DPANN) were combined into a protein database that was used in an hmmsearch against the arCOG database (settings: --tblout --domtblout --notextw)^{147,148}. Hits with e-values above 1e-3 were discarded from the domain results table, and the output was further parsed to only include the best hit for each protein based on the lowest e-value and highest bit score. The number of predicted genes that were assigned to an arCOG family is provided in Supplementary Data 1 for each genome. Note that in our focal set of 257 taxa, the proportion of predicted genes that can be assigned to an arCOG ranges from 47.08% to 99.8%, with a median of 82.3%.

We first identified and split potentially fused proteins comprising more than one protein domain, each assigned to a distinct arCOG (from now on called “splits”), using a custom pipeline to avoid incorrect inference of gene loss events (because each protein can only be assigned to its best arCOG hit). Proteins were further investigated for secondary domain hits after subtracting the position of the first domain hit from the full protein using bedtools subtract (v2.26.0)¹⁵⁸. Secondary domains were only considered if they were assigned to a different arCOG than the first domain. This process was repeated until proteins were investigated for up to four domains, and the original protein was split into individual domains using the positional domain information with bedtools getfasta. For reference, unsplit proteins were given the extension “a0” and the split proteins were given the extension “a1” to “a4,” indicating the primary to quarterly domain, in the generated trees.

All protein sequences were combined, and sequences with ambiguous amino acids were identified and discarded using a custom Python script (remove_seq_with_specific_char.py, <https://zenodo.org/records/18701544>). Afterward, a first round of alignments was generated using MAFFT or MAFFT-L-INS-i for gene families with more than or fewer than 1000 sequences, respectively (v7.407, settings: --any-symbol --reorder)¹⁴¹. Alignments were trimmed using trimAl (v1.4.rev15, settings: -gappypout)¹⁵⁹. Sequences with equal to or greater than 50% of gaps were identified and discarded using custom python scripts (faa_drop.py and screen_list_new.pl, <https://zenodo.org/records/18701544>). These steps allow to minimise the biases in phylogenetic inference induced by missing data^{160–162}. Only clusters with at least four sequences were considered for further analysis. Based on these initial selections, a second alignment was performed with MAFFT and MAFFT-L-INS-i as described above, and sequences were trimmed with BMGE (v1.12, settings: -t AA -m BLOSUM30 -b 2 -h 0.55)¹⁴². The trimmed alignments were used to generate a first round of single gene trees using IQ-TREE 2 (v2.1.2, settings: -m LG + G -T 1)⁶³. To minimise the

potential reconciliation noise introduced by the LBA artefacts in single gene trees, single sequences or clusters of sequences on long branches were identified using a custom script and using a branch length cutoff of two (cut_gene_tree.py, <https://zenodo.org/records/18701544>)¹⁵. The validity of this cutoff was confirmed on the basis of a random number of single gene trees that were manually inspected after adding the KO mapping, PFAM, and arCOG mappings as well as taxon information to the tip labels. Single sequences on long branches were discarded, and clusters of at least four sequences were assigned to a separate cluster using a custom script (screen_list_new.pl, <https://zenodo.org/records/18701544>). Finally, sequences within each cluster were subset to the focal taxon set of 257 archaeal genomes with a balanced representation of Euryarchaeota, TACK+Asgard and DPANN lineages, amenable to complex gene-tree and species-tree reconciliation models (<https://zenodo.org/records/18701544>).

A third round of alignments was generated with MAFFT and/or MAFFT-L-INS-i and trimmed with BMGE as described above. The trimmed alignment was used to perform a model test with IQ-TREE 2 (v2.1.2, settings: -m MF -mset LG -madd LG + C10, LG + C10 + G, LG + C10 + R, LG + C10 + F, LG + C10 + R + F, LG + C10 + G + F, LG + C20, LG + C20 + G, LG + C20 + F, LG + C20 + G + F, LG + C20 + R, LG + C20 + R + F --score-diff all) as well as to infer guide trees (settings: -m LG + G + F --score-diff all -nt 1 -wbtl -B 1000 -bnni). Depending on the best model chosen, single gene trees were inferred using either the posterior mean site frequency (PMSF) model (settings: -s trimmed_aln.faa -m best_model -ft guide_tree.treefile -T 1 -wbtl -B 1000 -pers 0.2 -nstop 500)¹⁶³ or the default model settings of IQ-TREE 2 (v2.1.2, settings: -s trimmed_aln.faa -m best_model -T 1 -wbtl -B 1000 -pers 0.2 -nstop 500) depending on whether or not a C-series profile mixture model was selected as the best model, respectively. In total, 7059 single gene trees were generated for the arCOG-based clusters, respectively, and used as input for the ancestral reconciliations as described below.

Gene tree-species tree reconciliation. To account for the incomplete nature of some MAGs, genome completeness estimates from CheckM v2 were used in reconciliation¹³⁹ (Supplementary Data 1). Bootstrapped single gene trees (1000 per gene family), inferred from 7,059 arCOG-based clusters (see arCOG gene families for ancestral gene content reconstruction), were reconciled with 15 different rooted archaeal species trees (Fig. 1A) using the distinct reconciliation models implemented in AleRax v1.1.1⁵⁶. These models include a global rate model (3 parameters, one D, one T and one L homogeneous to all branches), a per-family DTL rate model (21177 parameters, one D, one T and one L per gene family), and different branch-wise DTL models to capture differences in genome-evolving rates across archaeal lineages via the --model-parametrization option. All optimisations were performed using a limited-memory BFGS algorithm following recommendations from Williams et al.⁶⁵. We built a series of branch-heterogeneous gene family reconciliation models on top of a global-rate model as baseline (Supplementary Data 33). First, we introduced DPANN-specific rate classes: a different loss rate (DPANN:L; 4 parameters), different duplication+loss rates (DPANN:DL; 5 parameters), different transfer +loss rates (DPANN:TL; 5 parameters) and different duplication +transfer+loss rates (DPANN:DTL; 6 parameters). We then extended the DTL heterogeneity to major clades—Euryarchaeota, TACK+Asgard, DPANN and three DPANN sublineages (Woesearchaeales+SCGC-AAO11-G17, Pacearchaeales, Micrarchaeota), resulting in 21 parameters (DTL_br1). Building on this, we allowed clade-specific origination (O) rates for DPANN, Euryarchaeota and TACK+Asgard (25 parameters, DTL_br1_O). Next, we let D, T and L vary freely for Euryarchaeota lineages (i.e., for each Halobacteriota; Thermoplasmatota; and a clade comprising Methanobacteriota_A(B), Hydrothermarchaeota and Hadarchaeota), for TACK+Asgard lineages (i.e., for each Korarchaeota, Thermoproteota, Asgardarchaeota), and for DPANN lineages (i.e., for each Micrarchaeota; B1Sed10-29;

lainarchaeota; Altiarchaeota; Undinarchaeota; Aenigmarchaeota; PWEA01; EX4484_52; Nanohaloarchaeota; SpSt-1190; parvarchaeal-related lineages; UBA10117; Pacearchaeales; Woesearchaeales+SCGC-AAA011-G17; Nanoarchaeales; plus two internal nodes representing the last common ancestor of all DPANN and of DPANN Cluster 1 archaea, respectively), yielding 72 parameters in total (DTL_br2). Based on this model, in another experiment, we further allowed the D, T and L for Methanobacteriota_A, Methanobacteriota_B, Hydrothermarchaeota and Hadarchaeota to vary freely, resulting in 81 parameters. Model fit across models is assessed using the BIC and provided in Supplementary Data 10. The likelihoods of 7059 families under all of the above models were compared and tested using AU tests as implemented in CONSEL v0.20 with 1,000,000 bootstrap samples¹⁵⁶. In addition, we reanalysed data generated by Williams et al, 2017¹⁸ using a global model, a per-family DTL model, and a branch-wise model allowing Altiarchaeota, Micrarchaeota, lainarchaeota, Nanohaloarchaeota, Aenigmarchaeota, Parvarchaeota, Nanoarchaeales, Woesearchaeales, Pacearchaeales, Halobacteriota, Thermoplasmatota, Asgardarchaeota, Korarchaeota, Thermoproteota to evolve at different DTL rates, while we used different DTLO rates for nodes representing the LCA of DPANN, Euryarchaeota and TACK+Asgard, respectively (Supplementary Data 14 and 33). For ancestral genome content reconstruction, we allowed heterogeneous origination by estimating independent rates for the root, its two immediate descendants, and the DPANN lineage using --origination OPTIMISE, based on the reconciliation model with 72 parameters, which is highly correlated with the best-fitting model, DTL_br3 (Pearson's r : 1) (Supplementary Note 4).

Gene family filtration. To evaluate the underlying support for competitive root candidates, we performed gene family filtration similar to Coleman et al.¹². We ordered gene families by species number and taxonomic bias score (defined as $|1/3\%(\text{DPANN})| + |1/3\%(\text{TACK} + \text{Asgard})| + |1/3\%(\text{Euryarchaeota})|$), which quantifies the deviation from equal representation of three archaea groups in a gene family. To assess the potential of gene families to induce the SGA, we compared the differences in likelihoods between competing root hypotheses and the best-supported maximum-likelihood root as gene families were sequentially excluded from the analysis (Supplementary Note 2).

Ancestral sequence reconstruction and temperature estimates

To estimate the ancestral optimal growth temperature of LACA, we reconstructed ancestral sequences based on the supermatrix using the LG + G4+COaLA[2] model (best-fitting model from LG + G4+COaLA[1-3])¹⁰⁷ based on BIC; [x] refers to the number of independent axis positions per branch and on the root) implemented in Bio ++ (v2.3.1)¹⁶⁴. We sampled 100 sequences from the posterior distribution of ancestral states (i.e., amino acid) at each node to account for reconstruction uncertainty. Correspondence analyses on amino acid compositions were performed using the Prince¹⁶⁵ package (v0.13.0) Python3¹⁶⁶. Axis 2 of amino acid compositions was found to be significantly correlated with the optimal growth temperatures in extant archaeal taxa (Pearson's r : -0.87, p -value: 8.951×10^{-20} , Fig. S22). A linear regression model was fitted using the OLS function implemented in statsmodels (v0.13.5)¹⁶⁷, and the fitted (R^2 : 0.75) regression line was used to predict the OGT for ancestral nodes and extant taxa without known OGT. Confidence intervals were computed to account for uncertainty in the regression line parameter estimates.

Genome size estimation

The number of arCOG families present at the ancestral nodes of interest was obtained by counting families with posterior probabilities (PP) ≥ 0.5 (moderate support) and ≥ 0.75 (strong support). To estimate ancestral genome size, we used all 7059 gene families. Specifically, we applied locally weighted scatterplot smoothing (LOESS) to model the relationship between protein number and genome size across extant

genomes and predicted ancestral genome sizes from the fitted curve. Prediction uncertainty was quantified by bootstrap resampling (1000 replicates).

Statistical tests

Statistical significance of transfer propensity and gene duplication, loss and transfer events of different groups was tested using a Kruskal–Wallis test, a one-sided statistical test implemented in scipy-stats (v1.16.2), followed by a post hoc pairwise test for multiple comparisons of mean rank sums (Conover's test) using scikit_posthocs (v0.11.2) with P -values adjusted via the Holm–Bonferroni method. The correlation between ancestral state reconstructions inferred via distinct reconciliation models was computed using the corrcor function implemented in numpy (v1.26.4), and hierarchical clustering of the reconciliation models was performed using the linkage function implemented in scipy (v1.16.2) with “average” method.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All genomic data analysed are available at NCBI (Supplementary Data 1), and also deposited in the Zenodo repository (<https://zenodo.org/records/18701544>). Data generated in this study, including single gene tree analyses and concatenated phylogenies (i.e., sequence files, alignment tree files, inferred reconciliation models and events), are also deposited in the Zenodo repository (<https://zenodo.org/records/18701544>).

Code availability

The workflows for inferring phylogenies, performing reconciliation and ancestral gene content reconstructions, and custom scripts are deposited in the repository (<https://zenodo.org/records/18701544>).

References

1. Rinke, C. et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
2. Brown, C. T. et al. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).
3. Castelle, C. J. et al. Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr. Biol.* **25**, 690–701 (2015).
4. Hug, L. A. et al. A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
5. Spang, A. et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).
6. Zaremba-Niedzwiedzka, K. et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).
7. Spang, A., Mahendrarajah, T. A., Offre, P. & Stairs, C. W. Evolving perspective on the origin and diversification of cellular life and the virosphere. *Genome Biol. Evol.* **14**, evac034 (2022).
8. Göker, M. & Oren, A. Valid publication of names of two domains and seven kingdoms of prokaryotes. *Int. J. Syst. Evol. Microbiol.* **74**, <https://doi.org/10.1099/ijsem.0.006242> (2024).
9. Dutkiewicz, Z. et al. Proposal of gen. nov., sp. nov. in the ubiquitous bacterial phylum phyl. nov. *ISME Commun.* **5**, ycae147 (2025).
10. Nakajima, M. et al. *Minisyncoccus archaeiphilus* gen. nov., sp. nov., a mesophilic, obligate parasitic bacterium and proposal of *Minisyncoccaceae* fam. nov., *Minisyncoccales* ord. nov., *Minisyncoccia* class. nov. and *Minisyncocota* phyl. nov. formerly referred to as *Patescibacteria* or candidate phyla radiation. *Int. J. Syst. Evol. Microbiol.* **75**, 006668 (2025).

11. Taib, N. et al. Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. *Nat. Ecol. Evol.* **4**, 1661–1672 (2020).
12. Coleman, G. A. et al. A rooted phylogeny resolves early bacterial evolution. *Science* **372**, eabe0511 (2021).
13. Moody, E. R. R. et al. An estimate of the deepest branches of the tree of life from ancient vertically evolving genes. *Elife* **11**, e66695 (2022).
14. Martinez-Gutierrez, C. A. & Aylward, F. O. Phylogenetic signal, congruence, and uncertainty across bacteria and Archaea. *Mol. Biol. Evol.* **38**, 5514–5527 (2021).
15. Davin, A. A. et al. A geological timescale for bacterial evolution and oxygen adaptation. *Science* **388**, eadp1853 (2025).
16. Mahendrarajah, T. A. & Spang, A. DPANN Archaea and CPR Bacteria: insights into early cellular evolution?. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **380**, 20240096 (2025).
17. Baker, B. A. et al. Phylogenomic analyses indicate the archaeal superphylum DPANN originated from free-living euryarchaeal-like ancestors. *Nat. Microbiol.* **10**, 1593–1604 (2025).
18. Williams, T. A. et al. Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl. Acad. Sci. USA* **114**, E4602–E4611 (2017).
19. Aouad, M. et al. 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).
20. Raymann, K., Brochier-Armanet, C. & Gribaldo, S. The two-domain tree of life is linked to a new root for the Archaea. *Proc. Natl. Acad. Sci. USA* **112**, 6670–6675 (2015).
21. Guy, L. & Ettema, T. J. G. The archaeal ‘TACK’ superphylum and the origin of eukaryotes. *Trends Microbiol.* **19**, 580–587 (2011).
22. Tamarit, D. et al. Description of *Asgardarchaeum abyssi* gen. nov. spec. nov., a novel species within the class Asgardarchaeia and phylum Asgardarchaeota in accordance with the SeqCode. *Syst. Appl. Microbiol.* **47**, 126525 (2024).
23. Imachi, H. et al. *Promethearchaeum syntrophicum* gen. nov., sp. nov., an anaerobic, obligately syntrophic archaeon, the first isolate of the lineage ‘Asgard’ archaea, and proposal of the new archaeal phylum Promethearchaeota phyl. Nov. and Promethearchaeati kingdom regn. nov. *Int. J. Syst. Evol. Microbiol.* **74**, 006435 (2024).
24. Castelle, C. J. & Banfield, J. F. Major new microbial groups expand diversity and alter our understanding of the tree of life. *Cell* **172**, 1181–1197 (2018).
25. Castelle, C. J. et al. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat. Rev. Microbiol.* **16**, 629–645 (2018).
26. Dombrowski, N., Lee, J.-H., Williams, T. A., Offre, P. & Spang, A. Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS Microbiol. Lett.* **366**, <https://doi.org/10.1093/femsle/fnz008> (2019).
27. Probst, A. J. et al. Biology of a widespread uncultivated archaeon that contributes to carbon fixation in the subsurface. *Nat. Commun.* **5**, 5497 (2014).
28. Youssef, N. H. et al. Insights into the metabolism, lifestyle and putative evolutionary history of the novel archaeal phylum ‘Diapherotrites’. *ISME J.* **9**, 447–460 (2015).
29. Huber, H. et al. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* **417**, 63–67 (2002).
30. Hamm, J. N. et al. Unexpected host dependency of Antarctic Nanohaloarchaeota. *Proc. Natl. Acad. Sci. USA* **116**, 14661–14670 (2019).
31. La Cono, V. et al. Symbiosis between nanohaloarchaeon and haloarchaeon is based on utilization of different polysaccharides. *Proc. Natl. Acad. Sci. USA* **117**, 20223–20234 (2020).
32. Sakai, H. D. et al. Insight into the symbiotic lifestyle of DPANN archaea revealed by cultivation and genome analyses. *Proc. Natl. Acad. Sci. USA* **119**, e2115449119 (2022).
33. Krause, S. et al. The importance of biofilm formation for cultivation of a Micrarchaeon and its interactions with its Thermoplasmatales host. *Nat. Commun.* **13**, 1735 (2022).
34. Wurch, L. et al. Genomics-informed isolation and characterization of a symbiotic Nanoarchaeota system from a terrestrial geothermal environment. *Nat. Commun.* **7**, 12115 (2016).
35. St John, E. et al. A new symbiotic nanoarchaeote (*Candidatus Nanocleptia minutus*) and its host (*Zestosphaera tikiterensis* gen. nov., sp. nov.) from a New Zealand hot spring. *Syst. Appl. Microbiol.* **42**, 94–106 (2019).
36. Kato, S. et al. *Nanobdella aerobiophila* gen. nov., sp. nov., a thermoacidophilic, obligate ectosymbiotic archaeon, and proposal of *Nanobdellaceae* fam. nov., *Nanobdellales* ord. nov. and *Nanobdellia* class. nov. *Int. J. Syst. Evol. Microbiol.* **72**, <https://doi.org/10.1099/ijsem.0.005489> (2022).
37. Huang, W.-C. & Spang, A. DPANN archaea. *Curr. Biol.* **35**, R791–R794 (2025).
38. Reysenbach, A.-L. et al. Complex subsurface hydrothermal fluid mixing at a submarine arc volcano supports distinct and highly diverse microbial communities. *Proc. Natl. Acad. Sci. USA* **117**, 32627–32638 (2020).
39. Mei, R., Kaneko, M., Imachi, H. & Nobu, M. K. The origin and evolution of methanogenesis and are intertwined. *Proc. Natl. Acad. Sci. Nexus* **2**, gad023 (2023).
40. Thauer, R. K., Kaster, A.-K., Seedorf, H., Buckel, W. & Hedderich, R. Methanogenic archaea: ecologically relevant differences in energy conservation. *Nat. Rev. Microbiol.* **6**, 579–591 (2008).
41. Evans, P. N. et al. An evolving view of methane metabolism in the Archaea. *Nat. Rev. Microbiol.* **17**, 219–232 (2019).
42. Aouad, M. et al. Extreme halophilic archaea derive from two distinct methanogen Class II lineages. *Mol. Phylogenet. Evol.* **127**, 46–54 (2018).
43. Feng, Y. et al. The evolutionary origins of extreme halophilic archaeal lineages. *Genome Biol. Evol.* **13**, evab166 (2021).
44. Brochier, C., Gribaldo, S., Zivanovic, Y., Confalonieri, F. & Forterre, P. Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales?. *Genome Biol.* **6**, R42 (2005).
45. Petitjean, C., Deschamps, P., López-García, P. & Moreira, D. Rooting the domain archaea by phylogenomic analysis supports the foundation of the new kingdom Proteoarchaeota. *Genome Biol. Evol.* **7**, 191–204 (2014).
46. Wang, B. et al. Expansion of Thaumarchaeota habitat range is correlated with horizontal transfer of ATPase operons. *ISME J.* **13**, 3067–3079 (2019).
47. Mahendrarajah, T. A. et al. ATP synthase evolution on a cross-braced dated tree of life. *Nat. Commun.* **14**, 7456 (2023).
48. Baker, B. A. et al. Expanded phylogeny of extremely halophilic archaea shows multiple independent adaptations to hypersaline environments. *Nat. Microbiol.* **9**, 964–975 (2024).
49. Tarrio, R., Rodríguez-Trelles, F. & Ayala, F. J. Tree rooting with outgroups when they differ in their nucleotide composition from the ingroup: the *Drosophila* saltans and willistoni groups, a case study. *Mol. Phylogenet. Evol.* **16**, 344–349 (2000).
50. Gouy, R., Baurain, D. & Philippe, H. Rooting the tree of life: the phylogenetic jury is still out. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140329 (2015).
51. Naser-Khdour, S., Quang Minh, B. & Lanfear, R. Assessing confidence in root placement on phylogenies: an empirical study using nonreversible models for mammals. *Syst. Biol.* **71**, 959–972 (2022).

52. Squartini, F. & Arndt, P. F. Quantifying the stationarity and time reversibility of the nucleotide substitution process. *Mol. Biol. Evol.* **25**, 2525–2535 (2008).
53. Yap, V. B. & Speed, T. Rooting a phylogenetic tree with non-reversible substitution models. *BMC Evol. Biol.* **5**, 2 (2005).
54. Yang, Z. & Roberts, D. On the use of nucleic acid sequences to infer early branchings in the tree of life. *Mol. Biol. Evol.* **12**, 451–458 (1995).
55. Galtier, N. & Gouy, M. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.* **15**, 871–879 (1998).
56. Morel, B., Williams, T. A., Stamatakis, A. & Szöllösi, G. J. AleRax: a tool for gene and species tree co-estimation and reconciliation under a probabilistic model of gene duplication, transfer, and loss. *Bioinformatics* **40**, btac162 (2024).
57. Szöllösi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E. & Daubin, V. Efficient exploration of the space of reconciled gene trees. *Syst. Biol.* **62**, 901–912 (2013).
58. Muñoz-Gómez, S. A. et al. Site-and-branch-heterogeneous analyses of an expanded dataset favour mitochondria as sister to known Alphaproteobacteria. *Nat. Ecol. Evol.* **6**, 253–262 (2022).
59. McCarthy, C. G. P., Susko, E. & Roger, A. J. Modeling site-and-branch-heterogeneity with GFMix. *bioRxiv* <https://doi.org/10.1101/2025.08.07.669136> (2025).
60. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
61. Dombrowski, N. et al. Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. *Nat. Commun.* **11**, 3939 (2020).
62. Moody, E. R. R. et al. The nature of the last universal common ancestor and its impact on the early Earth system. *Nat. Ecol. Evol.* **8**, 1654–1666 (2024).
63. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
64. Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508 (2002).
65. Williams, T. A. et al. Phylogenetic reconciliation: making the most of genomes to understand microbial ecology and evolution. *ISME J.* **18**, wrac129 (2024).
66. Wilkinson, M., McInerney, J. O., Hirt, R. P., Foster, P. G. & Embley, T. M. Of clades and clans: terms for phylogenetic relationships in unrooted trees. *Trends Ecol. Evol.* **22**, 114–115 (2007).
67. Shimodaira, H. & Hasegawa, M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114–1116 (1999).
68. Zhang, C., Nielsen, R. & Mirarab, S. ASTER: a package for large-scale phylogenomic reconstructions. *Mol. Biol. Evol.* **42**, msaf172 (2025).
69. Wong, T. K. F. et al. MAST: Phylogenetic inference with mixtures across sites and trees. *Syst. Biol.* **73**, 375–391 (2024).
70. Wu, Z., Liu, S. & Ni, J. Metagenomic characterization of viruses and mobile genetic elements associated with the DPANN archaeal superphylum. *Nat. Microbiol.* **9**, 3362–3375 (2024).
71. Cloarec, L. A. et al. Lineage-dependent partitioning of activities in chemoclines defines Woesearchaeota ecotypes in an extreme aquatic ecosystem. *Microbiome* **12**, 249 (2024).
72. Vázquez-Campos, X. et al. Genomic insights into the inhabiting an Australian radioactive legacy site. *Front. Microbiol.* **12**, 732575 (2021).
73. Rao, Y.-Z. et al. Horizontal gene transfer and gene loss drove the divergent evolution of host dependency in Micrarchaeota. *Nat. Sci. Rev.* **13**, nwaf542 (2025).
74. Dang, C. C. et al. nQMaker: estimating time nonreversible amino acid substitution models. *Syst. Biol.* **71**, 1110–1123 (2022).
75. Hamm, J. N. et al. New lineages provide insights into the convergent evolution of extreme salt adaptation within symbiotic Archaea. *Mol. Biol. Evol.* **43**, msag091 (2026).
76. Eme, L. et al. Inference and reconstruction of the heimdallarchaeal ancestry of eukaryotes. *Nature* **618**, 992–999 (2023).
77. Williams, T. A. et al. Inferring the deep past from molecular data. *Genome Biol. Evol.* **13**, evab067 (2021).
78. Huang, W.-C. et al. Phylogenomic analyses reveal that Pan-guarchaeum is a clade of genome-reduced Asgard Archaea within the Njordarchaeia. *Mol. Biol. Evol.* **42**, msaf201 (2025).
79. Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. USA* **96**, 3801–3806 (1999).
80. Tria, F. D. K. & Martin, W. F. Gene duplications are at least 50 times less frequent than gene transfers in prokaryotic genomes. *Genome Biol. Evol.* **13**, evab224 (2021).
81. Treangen, T. J. & Rocha, E. P. C. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.* **7**, e1001284 (2011).
82. Nelson-Sathi, S. et al. Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc. Natl. Acad. Sci. USA* **109**, 20537–20542 (2012).
83. Zerulla, K. & Soppa, J. Polyploidy in haloarchaea: advantages for growth and survival. *Front. Microbiol.* **5**, 274 (2014).
84. Raymann, K., Forterre, P., Brochier-Armanet, C. & Gribaldo, S. Global phylogenomic analysis disentangles the complex evolutionary history of DNA replication in archaea. *Genome Biol. Evol.* **6**, 192–212 (2014).
85. Makarova, K. S. & Koonin, E. V. Archaeology of eukaryotic DNA replication. *Cold Spring Harb. Perspect. Biol.* **5**, a012963 (2013).
86. Makarova, K. S., Sorokin, A. V., Novichkov, P. S., Wolf, Y. I. & Koonin, E. V. Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol. Direct* **2**, 33 (2007).
87. Albers, S.-V. & Jarrell, K. F. The archaeellum: how Archaea swim. *Front. Microbiol.* **6**, 23 (2015).
88. Huber, H. et al. A dicarboxylate/4-hydroxybutyrate autotrophic carbon assimilation cycle in the hyperthermophilic Archaeum *Ignicoccus hospitalis*. *Proc. Natl. Acad. Sci. USA* **105**, 7851–7856 (2008).
89. Loder, A. J. et al. Reaction kinetic analysis of the 3-hydroxypropionate/4-hydroxybutyrate CO fixation cycle in extremely thermoacidophilic archaea. *Metab. Eng.* **38**, 446–463 (2016).
90. Berg, I. A. et al. Autotrophic carbon fixation in archaea. *Nat. Rev. Microbiol.* **8**, 447–460 (2010).
91. Garritano, A. N., Song, W. & Thomas, T. Carbon fixation pathways across the bacterial and archaeal tree of life. *Proc. Natl. Acad. Sci. Nexus* **1**, gac226 (2022).
92. Fuchs, G. Alternative pathways of carbon dioxide fixation: insights into the early evolution of life?. *Annu. Rev. Microbiol.* **65**, 631–658 (2011).
93. Adam, P. S., Borrel, G. & Gribaldo, S. Evolutionary history of carbon monoxide dehydrogenase/acetyl-CoA synthase, one of the oldest enzymatic complexes. *Proc. Natl. Acad. Sci. USA* **115**, E1166–E1173 (2018).
94. Borrel, G., Adam, P. S. & Gribaldo, S. Methanogenesis and the Wood-Ljungdahl pathway: an ancient, versatile, and fragile association. *Genome Biol. Evol.* **8**, 1706–1711 (2016).
95. Weiss, M. C. et al. The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.* **1**, 16116 (2016).
96. Sousa, F. L. & Martin, W. F. Biochemical fossils of the ancient transition from geoenergetics to bioenergetics in prokaryotic one

- carbon compound metabolism. *Biochim. Biophys. Acta* **1837**, 964–981 (2014).
97. Ueno, Y., Yamada, K., Yoshida, N., Maruyama, S. & Isozaki, Y. Evidence from fluid inclusions for microbial methanogenesis in the early Archaean era. *Nature* **440**, 516–519 (2006).
 98. Hippler, B. & Thauer, R. K. The energy conserving methyltetrahydromethanopterin:coenzyme M methyltransferase complex from methanogenic archaea: function of the subunit MtrH. *FEBS Lett.* **449**, 165–168 (1999).
 99. Adam, P. S., Kolyfetis, G. E., Bornemann, T. L. V., Vorgias, C. E. & Probst, A. J. Genomic remnants of ancestral methanogenesis and hydrogenotrophy in Archaea drive anaerobic carbon cycling. *Sci. Adv.* **8**, eabm9651 (2022).
 100. Adam, P. S., Borrel, G. & Gribaldo, S. An archaeal origin of the Wood-Ljungdahl HMPT branch and the emergence of bacterial methylotrophy. *Nat. Microbiol.* **4**, 2155–2163 (2019).
 101. Moody, E. R. R. et al. The emergence of metabolisms through Earth history and implications for biospheric evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **380**, 20240097 (2025).
 102. Søndergaard, D., Pedersen, C. N. S. & Greening, C. HydDB: A web tool for hydrogenase classification and analysis. *Sci. Rep.* **6**, 34212 (2016).
 103. Greening, C. et al. Genomic and metagenomic surveys of hydrogenase distribution indicate H₂ is a widely utilised energy source for microbial growth and survival. *ISME J.* **10**, 761–777 (2016).
 104. Garcia, P. S., De Anda, V., Baker, B. J., Gribaldo, S. & Borrel, G. Evolution and diversity of oxidoreductases involved in redox balance and energy conservation. *Nat. Ecol. Evol.* **10**, 550–565 (2026).
 105. Garcia, P. S., Gribaldo, S. & Borrel, G. Diversity and evolution of methane-related pathways in Archaea. *Annu. Rev. Microbiol.* **76**, 727–755 (2022).
 106. Spang, A. et al. 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).
 107. Groussin, M., Boussau, B. & Gouy, M. A branch-heterogeneous model of protein evolution for efficient inference of ancestral sequences. *Syst. Biol.* **62**, 523–538 (2013).
 108. Probst, A. J. & Moissl-Eichinger, C. Altiaarchaeales: uncultivated archaea from the subsurface. *Life* **5**, 1381–1395 (2015).
 109. Waters, E. et al. The genome of Nanoarchaeum equitans: insights into early archaeal evolution and derived parasitism. *Proc. Natl. Acad. Sci. USA* **100**, 12984–12988 (2003).
 110. Jahn, U. et al. Nanoarchaeum equitans and Ignicoccus hospitalis: new insights into a unique, intimate association of two archaea. *J. Bacteriol.* **190**, 1743–1750 (2008).
 111. Zhang, I. H. et al. Uncultivated DPANN archaea are ubiquitous inhabitants of global oxygen-deficient zones with diverse metabolic potential. *mBio* **15**, e0291823 (2024).
 112. Huang, W.-C. et al. Comparative genomic analysis reveals metabolic flexibility of Woese archaeota. *Nat. Commun.* **12**, 5281 (2021).
 113. Gaisin, V. A., van Wolferen, M., Albers, S.-V. & Pilhofer, M. Distinct life cycle stages of an ectosymbiotic DPANN archaeon. *ISME J* **18**, wrae076 (2024).
 114. Kato, S. et al. Cell surface architecture of the cultivated DPANN archaeon. *J. Bacteriol.* **206**, e0035123 (2024).
 115. Castelle, C. J. et al. Protein family content uncovers lineage relationships and bacterial pathway maintenance mechanisms in DPANN Archaea. *Front. Microbiol.* **12**, 660052 (2021).
 116. Li, L. et al. Comparative genomics provides insights into the genetic diversity and evolution of the DPANN superphylum. *mSystems* **6**, e0060221 (2021).
 117. Liu, Y. et al. Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature* **593**, 553–557 (2021).
 118. Zhang, J. et al. Deep origin of eukaryotes outside Heimdallarchaeia within Asgardarchaeota. *Nature* **642**, 990–998 (2025).
 119. Baker, B. J. et al. Diversity, ecology and evolution of Archaea. *Nat. Microbiol.* **5**, 887–900 (2020).
 120. Foster, P. G., Cox, C. J. & Embley, T. M. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 2197–2207 (2009).
 121. Williams, T. A., Cox, C. J., Foster, P. G., Szöllösi, G. J. & Embley, T. M. Phylogenomics provides robust support for a two-domains tree of life. *Nat. Ecol. Evol.* **4**, 138–147 (2020).
 122. Kivenson, V. et al. An archaeal genetic code with all TAG codons as pyrrolysine. *Science* **390**, eadu2404 (2025).
 123. Spang, A., Caceres, E. F. & Ettema, T. J. G. Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. *Science* **357**, eaaf3883 (2017).
 124. Elkins, J. G. et al. A korarchaeal genome reveals insights into the evolution of the Archaea. *Proc. Natl. Acad. Sci. USA* **105**, 8102–8107 (2008).
 125. Pan, J. et al. Wood-Ljungdahl pathway found in novel marine Korarchaeota groups illuminates their evolutionary history. *mSystems* **8**, e0030523 (2023).
 126. Tahon, G. et al. Phylogenomics and ancestral reconstruction of Korarchaeota reveals genomic adaptation to habitat switching. *bioRxiv* <https://doi.org/10.1101/2023.09.28.559970> (2023).
 127. Köstlbacher, S., Panagiotou, K., Tamarit, D. & Ettema, T. J. G. WitChi: Efficient detection and pruning of compositional bias in phylogenomic alignments using empirical chi-squared testing. *bioRxiv* <https://doi.org/10.1101/2025.07.14.663642> (2025).
 128. Parks, D. H. et al. GTDB release 10: a complete and systematic taxonomy for 715,230 bacterial and 17,245 archaeal genomes. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkaf1040> (2025).
 129. Szöllösi, G. J., Davin, A. A., Tannier, E., Daubin, V. & Boussau, B. Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140335 (2015).
 130. Werner, F. & Grohmann, D. Evolution of multisubunit RNA polymerases in the three domains of life. *Nat. Rev. Microbiol.* **9**, 85–98 (2011).
 131. Catchpole, R. J. & Forterre, P. The evolution of reverse gyrase suggests a nonhyperthermophilic last universal common ancestor. *Mol. Biol. Evol.* **36**, 2737–2747 (2019).
 132. Wang, Y. et al. A methylotrophic origin of methanogenesis and early divergence of anaerobic multicarbon alkane metabolism. *Sci. Adv.* **7**, eabj1453 (2021).
 133. Baross, J. A. & Hoffman, S. E. Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. *Orig. Life Evol. Biosph.* **15**, 327–345 (1985).
 134. Chyba, C. F., Thomas, P. J., Brookshaw, L. & Sagan, C. Cometary delivery of organic molecules to the early Earth. *Science* **249**, 366–373 (1990).
 135. McCutcheon, J. P. & Moran, N. A. Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* **10**, 13–26 (2011).
 136. Johnson, M. D. et al. Large attachment organelle mediates interaction between Nanobdellota archaeon YN1 and its host. *ISME J* **18**, wrae154 (2024).
 137. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
 138. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).
 139. Chklovskii, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).

140. Darling, A. E. et al. PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2**, e243 (2014).
141. Katoh, K., Misawa, K., Kuma, K.-I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
142. Criscuolo, A. & Gribaldo, S. 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).
143. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
144. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373 (2003).
145. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
146. Xie, R. et al. Expanding Asgard members in the domain of Archaea sheds new light on the origin of eukaryotes. *Sci. China Life Sci.* **65**, 818–829 (2022).
147. Tatusov, R. L. et al. The COG database: an updated version includes eukaryotes. *BMC Bioinforma.* **4**, 41 (2003).
148. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Archaeal Clusters of Orthologous Genes (arCOGs): an update and application for analysis of shared features between thermococcales, methanococcales, and methanobacteriales. *Life* **5**, 818–840 (2015).
149. Aramaki, T. et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
150. Bateman, A. et al. The Pfam protein families database. *Nucleic Acids Res.* **32**, D138–D141 (2004).
151. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
152. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermini, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
153. Yu, G., Lam, T. T.-Y., Zhu, H. & Guan, Y. Two methods for mapping and visualizing associated data on phylogeny using Ggtree. *Mol. Biol. Evol.* **35**, 3041–3043 (2018).
154. Sand, A. et al. tqDist: a library for computing the quartet and triplet distances between binary or general trees. *Bioinformatics* **30**, 2079–2080 (2014).
155. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
156. Shimodaira, H. & Hasegawa, M. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**, 1246–1247 (2001).
157. Wong, T. et al. IQ-TREE 3: Phylogenomic inference software using complex evolutionary models. *EcoEvoRxiv* <https://doi.org/10.32942/x2p62n> (2025).
158. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
159. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
160. Saurabh, K., Holland, B. R., Gibb, G. C. & Penny, D. Gaps: an elusive source of phylogenetic information. *Syst. Biol.* **61**, 1075–1082 (2012).
161. Lemmon, A. R., Brown, J. M., Stanger-Hall, K. & Lemmon, E. M. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. *Syst. Biol.* **58**, 130–145 (2009).
162. Roure, B., Baurain, D. & Philippe, H. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Mol. Biol. Evol.* **30**, 197–214 (2013).
163. Wang, H.-C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Syst. Biol.* **67**, 216–235 (2018).
164. Guéguen, L. et al. Bio++: efficient extensible libraries and tools for computational molecular evolution. *Mol. Biol. Evol.* **30**, 1745–1750 (2013).
165. Halford, M. Prince. GitHub <https://github.com/MaxHalford/prince> (2018).
166. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual: (Python Documentation Manual Part 2)*. (CreateSpace, 2009).
167. Seabold, S. & Perktold, J. statsmodels: Econometric and statistical modeling with Python. in *9th Python in Science Conference* <https://doi.org/10.25080/Majors-92bf1922-011> (2010).

Author contributions

A. Sp. conceived the study. A. Sp., N.D. and W.H. designed experiments and assembled the initial and focal datasets, respectively. N.D. performed annotations and developed the phylogenetic analysis workflow. N.D., T.A.M. and W.H. performed the phylogenetic analyses and reconciliations. W.H. performed ancestral reconstructions and rooting analyses. N.W., A. St., T.A.W. and G.J. Sz. contributed novel methods and software development. W.H., N.D., A. Sp., T.A.M., T.A.W. and G.J. Sz. analysed and interpreted the data. W.H., A.Sp. and T.A.W. wrote, and all authors edited and approved the manuscript.

Funding

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.Sp. and grant agreement No. 714774, GENECLOCKS to G.J.S.). A.St. and N.W. are funded by the EU ERA Chair (HORIZON-WIDERA-2022-TALENTS-01: 2023-2028) programme under project ID 101087081 (Comp-Biodiv-GR). Our research was also funded by the John Templeton Foundation (63451 to L.Sz., G.J.Sz., T.A.W. and A. Sp.). The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

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/s41467-026-76386-7>.

Correspondence and requests for materials should be addressed to Gergely J. Szöllősi, Tom A. Williams or Anja Spang.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026

¹Department of Marine Microbiology and Biogeochemistry, Royal Netherlands Institute for Sea Research (NIOZ), Den Burg, The Netherlands. ²Department of Evolutionary & Population Biology, Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, Amsterdam, The Netherlands.

³Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, Amsterdam, The Netherlands. ⁴Biodiversity Computing Group, Institute of Computer Science, Foundation for Research and Technology Hellas, Heraklion, Greece. ⁵Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany. ⁶Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany. ⁷Institute of Evolution, HUN-REN Center for Ecological Research, Budapest, Hungary. ⁸Model-Based Evolutionary Genomics Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan. ⁹Centre for Evolution, Department of Life Sciences, University of Bath, Bath, UK.

✉ e-mail: gergely.szollosi@oist.jp; tw2316@bath.ac.uk; anja.spang@nioz.nl