ISME

The ISME Journal, 2024, **18(1)**, wrae129

https://doi.org/10.1093/ismejo/wrae129

Advance access publication: 13 July 2024

Review Article

Phylogenetic reconciliation: making the most of genomes to understand microbial ecology and evolution

Tom A. Williams[1,*], Adrian A. Davin[2], Lénárd L. Szánthó[3,4], Alexandros Stamatakis[5,6,7], Noah A. Wahl[5], Ben J. Woodcroft[8], Rochelle M. Soo[9], Laura Eme[10], Paul O. Sheridan[11], Cecile Gubry-Rangin[12], Anja Spang[13,14], Philip Hugenholtz[9], Gergely J. Szöllősi[3,4,15,*]

[1]School of Biological Sciences, University of Bristol, Bristol BS81TQ, United Kingdom
[2]Department of Biological Sciences, Graduate School of Science, The University of Tokyo, 113-0033 Tokyo, Japan
[3]MTA-ELTE "Lendület" Evolutionary Genomics Research Group, Eötvös University, 1117 Budapest, Hungary
[4]Model-Based Evolutionary Genomics Unit, Okinawa Institute of Science and Technology Graduate University, 904-0495 Okinawa, Japan
[5]Biodiversity Computing Group, Institute of Computer Science, Foundation for Research and Technology Hellas, 70013 Heraklion, Greece
[6]Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies, 69118 Heidelberg, Germany
[7]Institute of Theoretical Informatics, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany
[8]Centre for Microbiome Research, School of Biomedical Sciences, Queensland University of Technology (QUT), Translational Research Institute, Woolloongabba, QLD 4102, Australia
[9]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, QLD 4072, Australia
[10]Unité d'Ecologie, Systématique et Evolution, Université Paris-Saclay, 91190 Gif-sur-Yvette, France
[11]School of Biological and Chemical Sciences, University of Galway, Galway H91 TK33, Ireland
[12]School of Biological Sciences, University of Aberdeen, Aberdeen AB24 3FX, United Kingdom
[13]Department of Marine Microbiology and Biogeochemistry, NIOZ, Royal Netherlands Institute for Sea Research, PO Box 59, 1790 AB Den Burg, The Netherlands
[14]Department of Evolutionary & Population Biology, Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, Amsterdam, The Netherlands
[15]Institute of Evolution, HUN REN Centre for Ecological Research, 1121 Budapest, Hungary

[*]Corresponding authors: Tom Williams, School of Biological Sciences, University of Bristol, 24 Tyndall Ave, Bristol BS8 1TQ, United Kingdom. Email: tom.a.williams@bristol.ac.uk and Gergely J. Szöllősi, Okinawa Institute of Science and Technology, 1919-1 Tancha, Onna-son, Kunigami-gun, Okinawa 904-0495, Japan. Email: gergely-szollosi@oist.jp

Abstract

In recent years, phylogenetic reconciliation has emerged as a promising approach for studying microbial ecology and evolution. The core idea is to model how gene trees evolve along a species tree and to explain differences between them via evolutionary events including gene duplications, transfers, and losses. Here, we describe how phylogenetic reconciliation provides a natural framework for studying genome evolution and highlight recent applications including ancestral gene content inference, the rooting of species trees, and the insights into metabolic evolution and ecological transitions they yield. Reconciliation analyses have elucidated the evolution of diverse microbial lineages, from Chlamydiae to Asgard archaea, shedding light on ecological adaptation, host–microbe interactions, and symbiotic relationships. However, there are many opportunities for broader application of the approach in microbiology. Continuing improvements to make reconciliation models more realistic and scalable, and integration of ecological metadata such as habitat, pH, temperature, and oxygen use offer enormous potential for understanding the rich tapestry of microbial life.

Keywords: phylogenetics, gene tree–species tree reconciliation, microbial evolution, horizontal gene transfer

Introduction

Phylogenetic trees are important in microbial ecology, because to understand an ecosystem you need to understand its evolutionary history across multiple levels, ranging from genes to species to entire communities. For example, horizontal transfer of key genes between members of a community can have a large impact on organismal and community function as a whole [1]. The identification of such genes is often achieved via manual comparisons of species and gene trees to identify putative transfers of interest. Recently, new gene tree–species tree reconciliation methods including RANGER-DTL 2 [2], TALE [3], and AleRax [4] have emerged that take this approach to the next level, allowing thousands of genes to be compared with species trees.

These phylogenetic reconciliation methods have many applications in the study of microbial ecology and evolution, and indeed in biology more broadly, beyond the detection of horizontal gene transfer (HGT). For example, they can be used to infer more accurate species and gene trees and to root them. Furthermore, these inferences can be used to reconstruct ancestral gene content or to evaluate the evidence for coevolution between host and parasite or symbiont lineages. Reconciliation analyses have also been used to improve the reconstruction of ancestral protein sequences for use in evolutionary biochemistry and synthetic biology applications [5], and for studying whole-genome duplication in land plants [6].

As with any kind of data analysis, the accuracy of phylogenetic reconciliation depends on the algorithms used and the

Received: 6 May 2024. **Revised:** 1 July 2024. **Accepted:** 12 July 2024

assumptions they make about the processes of evolution [7]. There has been substantial progress in method development in recent years [8], including the development of methods that can accommodate uncertainty in the evolutionary history of single genes [9–11], incomplete lineage sorting [12], and that allow the inferred rates of gene duplication, transfer, and loss (DTL) to vary across the species tree [4], resulting in more sophisticated software packages for performing these analyses. Here, we first review how reconciliation methods work and argue that they provide a natural framework for modelling microbial genome evolution. We then review a body of recent work that illustrates how reconciliation methods can provide insight into microbial ecology and evolution.

Modelling microbial evolution

The application of phylogenetics to the study of microbial ecology and evolution has proven powerful but does not come without limitations. With the advent of high-throughput sequencing, single-gene alignments often contain more taxa than sites (aligned amino acid or nucleotide positions) and may contain too little information with which to confidently resolve sequence relationships. Even well-supported gene trees are statistical estimates of evolutionary history and are not guaranteed to be correct. In particular, trees depend on the model of sequence evolution used to infer them, which describes the rates of different amino acid or nucleotide substitutions over time (substitution model). Numerous models are available, and model choice is important because poorly fitting models can result in the inference of an incorrect tree with high statistical support [7]. A specific issue for microbial evolution is that existing models, many of which have been in use for decades, were inferred from small datasets predominantly comprising closely related animals and plants. Therefore, these models may not be representative of prokaryotic evolution [13]. Recent work has aimed to develop more appropriate substitution models for prokaryotes and microbes more broadly, with packages such as MAMMaL, EDCluster, and QMaker now available for estimating new models [13–15].

The importance of the substitution model in phylogenetics is exemplified by long-running debates in the literature about the deep structure of the tree of life, in which analyses of the same data using different models resulted in support for either two [16] or three [17] primary domains of life [18]. Best practice in phylogenetics is therefore to evaluate the fit of a range of substitution models to a given dataset and perform inference using the best-fitting model, as judged by statistical tests such as the Bayesian information criterion or the Akaike information criterion. However, better-fitting models are generally more complex and less computationally tractable, such that difficult decisions have to be made about the trade-off between dataset size and model adequacy—the degree to which a model captures the evolutionary process that gave rise to the sequence data. The scalability of phylogenetic methods, both simple and complex, is an increasing challenge in microbial evolution research that aims to integrate the wealth of new genome data being generated by cultivation-free approaches.

HGT, an important driver of microbial evolution [19], presents another modelling challenge. On generation-to-generation timescales, inheritance in prokaryotes is usually vertical, with transmission of the genome from mother to daughter cells [20]. This genetic process gives rise to the species tree describing the relationships among lineages. Among closely related genomes, HGT and homologous recombination can sometimes act to homogenize the gene pool and reinforce species boundaries [21]. However, HGT over longer genetic distances induces differences between gene trees, and between the gene trees and the species tree. The cumulative effect of gene transfer is that very few, if any, prokaryotic genes share the same history as the lineages they reside in [22], with the possible exception of very young (i.e. recently evolved) genes that have not yet experienced transfer. When viewed on a long evolutionary timescale, lineages of prokaryotes (and, perhaps, microbial eukaryotes) might be regarded as "ships of Theseus", continuously remodelled by gene transfer despite clear continuity of inheritance from one generation to the next [23].

In addition to HGT, gene duplication and loss are fundamental processes that shape microbial genomes and need to be taken into account in any model. Duplicated genes are particularly common in eukaryotes [24] but are important for the emergence of novelty in all lifeforms, with duplicated genes experiencing less selective constraints that can enable the evolution of new functions [25]. Ancient gene duplications underpinned the evolution of core molecular complexes of prokaryotic cells, including the membrane-bound ATP synthase [26, 27] and the signal recognition particle/receptor system for targeting proteins to the cell membrane [28], but gene duplication also plays an important role in prokaryotes on more recent timescales. For example, exposure to antibiotics can promote the fixation of duplicate resistance genes [29] as a means of increasing gene dosage and therefore expression level [30], and gene duplication followed by functional divergence has been shown to drive the evolution of biosynthetic gene clusters that produce novel secondary metabolites in *Streptomyces* [31].

Gene loss is frequent in both prokaryotes and eukaryotes and underpins the evolution of symbiotic and parasitic lineages, including the nutritional endosymbionts of aphids [32], DPANN Archaea("DPANN" Archaea were originally defined as a superphylum containing Diapherotrites, Parvarchaeota, Aenigmaarchaeota, Nanoarchaeota, and Nanohaloarchaeota [33], but now also contain other small-genome lineages such as Woesearchaeota and Pacearchaeota [34].) [33, 34], Patescibacteria/CPR [35, 36], and parasitic fungi such as Microsporidia [37]. Although the lost genes often encode metabolic functions that are no longer required following the shift to a host-associated lifestyle, it has also been suggested that periods of gene loss may facilitate subsequent adaptive evolution via the disruption of preexisting gene interaction networks [38].

A complete picture of microbial genome evolution therefore requires consideration of HGT, gene duplication, and loss. These processes can be captured via a conceptual model of microbial evolution along the lines of that depicted in Fig. 1A, in which genes evolve vertically along a species tree, sometimes experiencing gene duplications, losses, and transfers into other contemporary lineages. The questions we might want to ask of this model include what is the overarching species tree? What are the relative contributions of vertical transmission and horizontal transfer to genome evolution, and do these vary over the tree? Where in evolutionary history did gene duplications, transfers, and losses occur? Which gene families were present at each internal node on the tree, and consequently which ancestral metabolic capabilities and environmental adaptations can be inferred at each time point?

The first of these questions—the topology of the species tree—can be addressed using concatenation or supertree approaches, whereby phylogenetic information is combined from multiple genes predicted to have been inherited vertically from a common ancestor (Fig. 1B). In the concatenation approach, initial phylogenetic analyses are used to identify a set of genes that, within

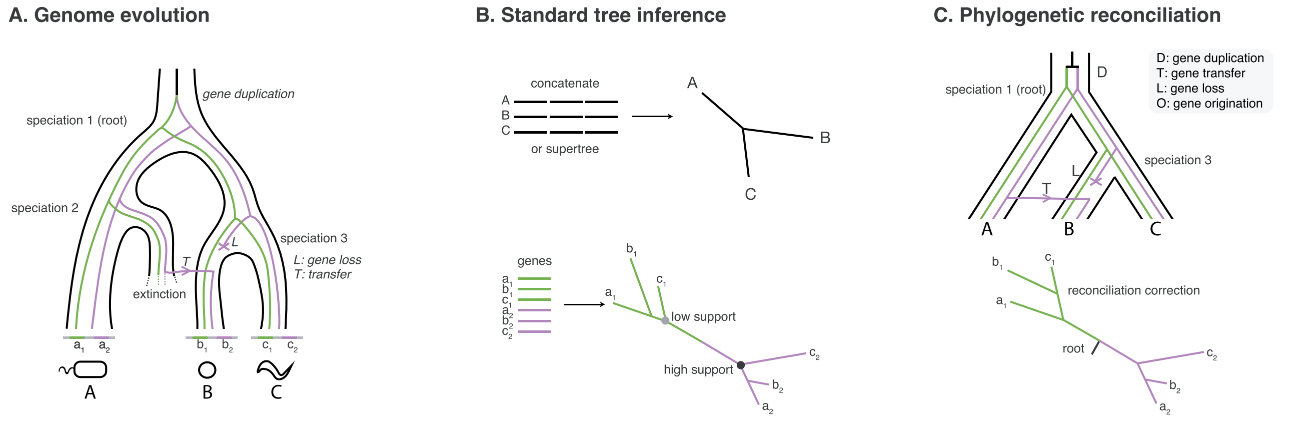

Figure 1. Modelling microbial evolution using phylogenetic reconciliation. (A) A conceptual model of microbial genome evolution, showing an overarching species (or lineage) tree with three speciations and one extinction event. Genes evolve along this tree, occasionally jumping between lineages by HGT (arrow), being duplicated (triangle), or being lost (cross). A, B, and C are different species; a_1 and a_2 are homologous genes that are part of the broader gene family being analysed, and which are both found in species A. (B) Standard phylogenetic inference reconstructs the tree from the sequence alignment using a stochastic substitution model. Single gene alignments are often short and contain limited information, so that the inferred gene tree can contain weakly supported branches. To estimate the species tree, information can be aggregated from multiple genes using concatenation or supertree approaches. However, these approaches estimate the topology but in most cases not the root of the species tree, unless an outgroup is included [117] or a nonreversible model is used [118, 119]. (C) To model microbial evolution, we need to capture not only the evolution of gene sequences along gene trees (via a substitution model) but also the evolution of gene trees along the species tree (via a reconciliation model that describes rates of DTL). The result is a kind of "species tree aware" phylogenetics, in which information from many gene families is used to infer the species tree, and in turn the histories of the individual gene families are contextualized as a series of gene duplication, loss, and transfer events. Phylogenetic reconciliation can combine this information by jointly optimising the phylogenetic (substitution) and reconciliation likelihoods. In the example gene tree in panel B, there is strong support for a sister relationship between the a_2 and b_2 genes (in the red clade), indicating a transfer event from Branch A to B on the species tree (Panel C); however, the sister relationship predicted for the a_1 and b_1 genes (in the blue clade) is weakly supported and likely incorrect, given the B + C relationship in the species tree. By incorporating information from the species tree, reconciled gene trees can show improved accuracy over the gene tree estimated from the sequence alignment alone [56].

the limits of statistical resolution, appear to have congruent evolutionary histories. Multiple sequence alignments for each of these genes are then stitched together and analysed as a single alignment, providing a simple way to pool phylogenetic information. User-friendly tools for inferring trees from sequence alignments are now available, with some of the most popular packages including RAxML-NG [39], IQ-TREE [40], and PhyloBayes [41]. By contrast, supertree or "species tree" methods take a two-step approach: single gene trees are inferred separately for each gene (using a program such as RAxML-NG or IQ-TREE) and then their phylogenetic signal is combined using one of a variety of heuristic algorithms [42–44], with ASTRAL and related tools among the most popular and performant methods [45]. An advantage of concatenation methods is that longer alignments allow the use of better-fitting but more complex substitution models, which have repeatedly been shown to be important for accurate inference of phylogenetic trees in deep time [46]. Supertree methods can be faster and are more robust to incomplete lineage sorting and HGT among the input gene trees than is concatenation [47]. It is encouraging that the results of concatenation and supertree analyses are often concordant and, in recent years, have been converging on the topology of the tree of life. For example, concatenation of core genes and supertree analysis of broadly shared gene families both recovered a "two domains" tree of life [48], and taxonomic schemes inferred from large-scale prokaryotic trees inferred using concatenation and supertree methods were 98.2% identical [49].

Beyond species tree estimation, answering the other questions about microbial genome evolution posed above requires comparison of gene and species trees. Manual comparison of individual gene trees with the species tree can identify putative gene transfers, but the approach lacks power because any given gene tree can be explained by many different combinations of DTL events. The problem is compounded by phylogenetic uncertainty, further increasing the number of possible scenarios, and by phylogenetic error. Phylogenetic reconciliation systematically addresses these issues by modelling the evolution of gene families in the shared context of the overarching species tree, capturing both sequence evolution and higher-level processes of DTL.

Figure 1C illustrates how the conceptual model of microbial evolution outlined above and in Fig. 1A can be operationalized using phylogenetic reconciliation methods. Reconciliations are mechanistically explicit scenarios that describe how a gene family has evolved on the species tree, starting with an origination event at a specific ancestral node, and followed by a series of events, each of which is mapped to a specific branch of the species tree. The reconciliation scenario ends with the members of the gene family observed in modern genomes arriving at the tips of the species tree. The set of possible events include vertical transmissions from ancestor to descendant node, gene duplications, transfers (from a donor to a recipient branch), and losses. For a given species tree and gene tree, many distinct reconciliation scenarios are possible, so we need criteria for choosing between them. One approach is to use the principle of parsimony: if we can assign relative costs to the different possible event types in advance (e.g. DTLs), we can find the reconciliation(s) that have the lowest summed cost [2, 50–53]. Alternatively, probabilistic model-based approaches can be used to estimate the rates of each type of event from the data using maximum likelihood or Bayesian methods [4, 9, 10, 54]. These are more computationally intensive but have the advantage that rates do not need to be set *a priori*. What we know of microbial ecology and evolution suggests that the relative rates of events vary across the tree of life. For example, duplications are more frequent in eukaryotes, transfers in prokaryotes, and losses in host-associated lineages [55]. As such, model-based approaches that estimate the DTL rates directly from the data have clear advantages for studying microbial evolution [56]. For example, reconciled gene trees inferred

Table 1. A selection of available reconciliation software that can model DTL; for a more comprehensive overview of available packages, see [8].

Package	Inference framework	Input	Remark	Repository	Reference
RANGER-DTL2	Parsimony	Single gene tree for each multiple sequence alignment (MSA)	Widely used parsimony reconciliation tool	https://compbio.engr.uconn.edu/software/RANGER-DTL/	[2]
TreeFix DTL	Parsimony	Single gene tree for each MSA	Widely used parsimony reconciliation tool	https://compbio.mit.edu/treefix-dtl/index.html#download	[11]
ecceTERA	Parsimony	Sample of gene trees for each MSA	Reconciliation of gene trees in a parsimony framework	https://mbb.univ-montp2.fr/MBB/subsection/softExec.php?soft=eccetera	[52]
ALE	Probabilistic (maximum likelihood, Bayesian)	Distribution of gene trees for each MSA	Calculate species tree likelihood and sample reconciled gene trees in a probabilistic framework	https://github.com/ssolo/ALE	[10]
TALE	Probabilistic (Bayesian)	Host species tree, single or distribution of symbiont and gene trees	Implements a three-level hierarchical reconciliation approach: genes within symbionts within hosts	https://github.com/hmenet/TALE	[3]
GeneRax	Probabilistic (maximum likelihood)	MSA	Directly estimates ML reconciled gene tree, good for very similar sequences	https://github.com/BenoitMorel/GeneRax	[59]
AleRax	Probabilistic (maximum likelihood)	Distribution of gene trees for each MSA	Calculate species tree likelihood and sample reconciled gene trees in a probabilistic framework	https://github.com/BenoitMorel/AleRax	[4]

using the probabilistic reconciliation method ALE (amalgamated likelihood estimation [10]) were more accurate than those inferred with a range of parsimony methods [56, 57]. Thus, while analyses with a diversity of reconciliation methods have provided insights into microbial ecology and evolution (see practical applications discussed in the next section and Table 1), our recommendation is that probabilistic model-based methods should be used where possible (see Box 1).

Box 1 Best practices in phylogenetic reconciliation.

The most common phylogenetic reconciliation task is mapping the evolutionary histories of a collection of gene families onto a rooted species tree. We describe one workflow using AleRax, a fast and accurate state-of-the-art tool recently developed by some of us, but we encourage interested users to explore the range of tools available. Each of the steps below is itself a large topic, and we refer the interested reader to further literature throughout. The software repository is usually the best resource for information on how to run tools. For example, see https://github.com/BenoitMorel/AleRax/wiki for AleRax or https://compbio.engr.uconn.edu/software/RANGER-DTL/ for RANGER-DTL2.

We will first infer the gene trees, and infer and root the species tree. We recommend using the most accurate methods for each of these steps, to the extent that time and computational resources allow. However, consider using faster, less accurate alignment and tree inference options or trading off dataset size when doing initial analyses, to save time and carbon emissions.

1. Gene tree inference.
   a. Work with the nucleotide or protein sequences from a set of genomes of interest and label each with the genome of origin and the unique gene ID. For example, "ECOLI_XBV38467." By default, AleRax will interpret what comes before the underscore as the species name, and what comes after as the gene ID, which will aid with mapping later.
   b. Cluster the sequences into gene families, either using established sets of homologous clusters (such as through COG or KEGG annotation of your set of genomes using eggNOG-mapper [123]), or a *de novo* gene clustering tool (such as mcl [124] or Broccoli [125]) Gene families must include all homologous genes, i.e. orthologues, paralogues, and xenologues.
   c. Align the gene families using e.g. muscle 5 [126].
   d. Infer trees for each gene family using software such as IQ-TREE2 [40], RAxML-NG [39], PhyloBayes [41], or MrBayes [127]. If using a maximum likelihood program, tell it to write the bootstrap trees to an output file (e.g. "-wbtl" in IQ-TREE2), because it is these—not the maximum likelihood tree—that capture the uncertainty about evolutionary relationships within the gene family. If using a Bayesian tool, use a sample of trees from the posterior distribution to represent this uncertainty. We recommend using a Bayesian tool if dataset size allows, because the conditional clade probabilities used in AleRax benefit from an accurate sample of the posterior distribution, which the maximum likelihood bootstrap can only approximate.
2. Species tree inference [47, 128].
   a. Use a best-practice approach to species tree inference—either using a concatenate of universal markers or supertree methods (see main text). There are several established tools for obtaining marker genes, including BUSCO [129], GTDB-Tk [130], or OrthoFinder [131].

    b. Root the species tree based on prior evidence or using an appropriate rooting method; e.g. in a study of opisthokonts, you might root the tree between Holozoa and Holomycota. If you are unsure of the root position, if inferring the root is the goal of your study, or if the root position is controversial, you might use the reconciliation analysis itself to infer the root (see below).

3. Gene tree–species tree reconciliation.
    a. Set up the AleRax gene families file, which specifies where the gene tree samples for each gene family are to be found (see AleRax wiki at https://github.com/BenoitMorel/AleRax/wiki/Running-AleRax#families-file).
    b. Run an AleRax analysis using the default model and a rooted fixed species tree, using 10 cores on a machine with MPI installed:

```
mpiexec -np 10 alerax -f families.txt -s rooted_species_tree.newick —species-tree-search SKIP.
```

Top tips

- For analyses of protein datasets on deep evolutionary timescales, we have found that deleting (or “trimming”) poorly aligned positions (e.g. using a tool such as BMGE or trimAl) can improve phylogenetic inference.
- When selecting marker genes for species tree inference, it is important to choose genes that are single copy in most of the species analysed. In the case of duplicate genes, mixing orthologues and paralogues will confound the phylogenetic signal.
- Two useful optional AleRax commands include –trim-ratio X and –memory-savings. Not to be confused with “trimming” alignment sites, –trim-ratio X will discard some proportion X of the families with the largest number of clades in their sample of input trees; this can greatly reduce total runtime. For example, –trim-ratio 0.05 will discard the 5% of families with the largest number of clades in the input tree sample. –memory-savings will substantially reduce RAM usage at the cost of 10-20% additional runtime, which can be useful when analysing large datasets or when using machines with limited RAM.
- Remove highly similar or identical sequences in gene family alignments.
- Explore using branch-wise DTL models if the gene evolutionary process is highly heterogeneous (cf. https://github.com/BenoitMorel/AleRax/wiki/Running-AleRax#model-parametrization).
- Although the standard optimizer (GRADIENT) is fast compared to other options, it might struggle occasionally to converge to a global optimum. If your analysis allows for more runtime, we recommend using the LBFGSB optimizer via the –rec-opt flag.

Probabilistic reconciliation-based approaches to studying microbial evolution

Probabilistic phylogenetic reconciliation can be considered a natural extension of traditional phylogenetic inference using substitution models. Just as the parameters of the substitution model include the gene tree and the relative rates of change between different nucleotide or amino acid states, the parameters of the reconciliation model describe the rooted species tree and the rates of DTL events. Reconciliation scenarios are then equivalent to substitution histories, with each sampled history representing a specific series of evolutionary events giving rise to the observed gene tree (or sequence alignment). In both cases, we sum over all possible scenarios when calculating the likelihood. As in the case of substitution models, we can choose among different reconciliation models that describe the evolutionary process in a more simple or complex way. For example, the simplest model in the phylogenetic reconciliation package AleRax [4] assumes a single set of DTL rate parameters for all genes and all branches of the species tree (Table 1). Alternatively, we might use a model in which rates vary among different clades or branches of the tree, potentially capturing important biological signal at the expense of additional model complexity and risk of overparameterization (i.e. the inclusion of additional unnecessary parameters that might increase computational cost and decrease accuracy).

The input to phylogenetic reconciliation packages varies (Table 1). Some methods (e.g. Phyldog [58], GeneRax [59]) take gene family sequence alignments as input and infer the rooted species tree as well as the reconciliations by jointly optimizing the likelihood of the reconciliation and substitution models. Most current methods use a two-step approach. First, gene trees are inferred using a standard phylogenetic reconstruction tool (perhaps IQ-TREE [40], RAxML-NG [39], or PhyloBayes [41]), which are then provided as input to the reconciliation software. Due to this inherent uncertainty of gene trees, some packages [4, 10] represent each gene family via a sample of plausible trees, obtained using bootstrapping or, in Bayesian analyses, by Markov Chain Monte Carlo sampling. Taking a sample of trees for each gene family captures this phylogenetic uncertainty, leading to reconciliations that better represent the potentially weak signal in the original sequence alignments [4]. The number of gene families reconciled in an analysis depends on the scientific question and can range from a single gene family to all available gene families on the set of genomes being analysed (see examples below).

How phylogenetic reconciliation has been applied in microbial ecology and evolution

The use of phylogenetic reconciliation methods in microbial ecology and evolution has recently gained popularity. Reconciliation-based approaches to ancestral gene content inference are particularly useful for studying microbial evolution because the reconstructions naturally incorporate phylogenetic evidence for HGT (Fig. 2). Probabilistic reconciliation methods have the additional benefit that they can accommodate phylogenetic uncertainty by averaging over possible reconciliation histories when inferring ancestral gene repertoires. In what follows, we briefly summarize some interesting applications that demonstrate how these methods can be used to link the tree of life and Earth history, reconstruct metabolic repertoires, infer past ecological transitions, and reconstruct the evolution of biogeochemical cycles (Fig. 3A–D). To date, most probabilistic reconciliation studies in microbial evolution have used ALE [10], which was the first efficient implementation of an algorithm that can account for gene tree uncertainty and model HGT. ALE has now been superseded by AleRax [4], which provides a faster, parallelizable and more flexible implementation of the original model, alongside other useful new features. In the examples below, we note when reconciliation packages other than ALE were used.

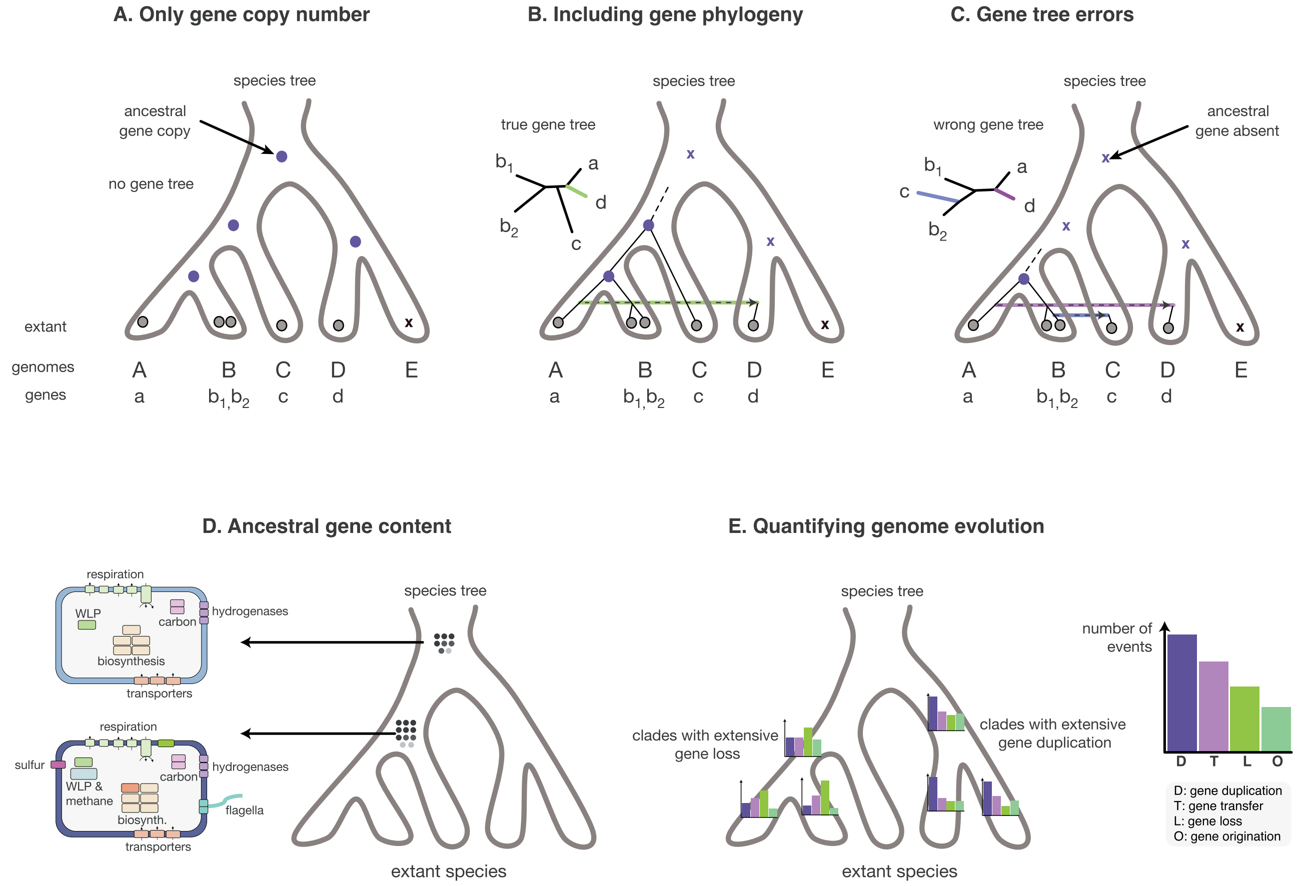

Figure 2. Reconstructing ancestral gene content using phylogenetic reconciliation. By "drawing" the gene tree into the species tree, phylogenetic reconciliation provides an explicit estimate of gene presence at internal nodes of the species tree and therefore ancestral gene complements. One advantage of reconciliation-based approaches to ancestral gene content inference is that the method accounts for gene transfer, as illustrated here. Grey dots denote observed gene copies in extant species, blue dots denote inferred ancestral presence of a gene, blue Xs denote inferred ancestral absence of a gene. (A) Consider a gene family broadly distributed in extant taxa. On the basis of this phylogenetic distribution alone, it appears likely that the gene traces to the root of the species tree. (B) However, a comparison of the gene tree to the species tree suggests a recent horizontal acquisition of the gene on the right hand side of the species tree root; as a result, the gene is inferred to have originated more recently. Comparison of (A) and (B) illustrates why methods based on gene copy number distribution alone may overestimate ancestral gene contents. (C) However, incorporating phylogenetic information comes at a potential cost: Errors in the reconstructed gene tree may be interpreted as additional gene transfers, so that the evolutionary age of the gene is underestimated. This case illustrates why reconciliation-based methods may tend to underestimate ancestral gene repertoires. (D) Performing reconciliation analyses for all gene families in a dataset results in the inference of gene contents at ancestral nodes, with per-family presence probabilities based upon the reconciliation model. By taking all gene families above a given probability threshold and cross-referencing with information about gene functions (e.g. based upon the COG or KEGG databases), ancestral metabolic capabilities can be inferred. (E) The reconstructed history of gene origination, DTL events on each branch of the species tree can be used to quantify genome dynamics through time, identifying periods of genome remodelling and evolutionary innovation. This figure is based on that of [120] with some modifications.

The origin of eukaryotes and the tree of life

Phylogenetic reconciliation has recently been used to reconstruct ancestral gene repertoires among the Asgard archaea, the archaeal lineage most closely related to eukaryotes [60]. The analysis indicated that rates of gene duplication were higher in two Asgard lineages, the Lokiarchaeales and Hodarchaeales, the latter of which appear to be the closest living archaeal relatives of the eukaryotic cell. This result suggests that some of the genome dynamics that distinguish prokaryotes and eukaryotes (such as a higher rate of gene duplication in eukaryotes) might have predated the prokaryote-to-eukaryote transition [60], also recently suggested by other analyses [61]. The analysis also suggested that the archaeal lineage ancestral to eukaryotes likely lost the capacity for autotrophic growth using genes of the Wood-Ljungdahl pathway prior to entering symbiosis with the bacterial ancestor of the mitochondrion, arguing against eukaryogenesis scenarios in which the archaeal partner was a hydrogen-dependent autotroph [60].

Looking still further back in evolutionary history, phylogenetic reconciliation was used to root the bacterial phylogeny (Fig. 3A and B) [56]. The advantage of this approach is that the bacterial tree could be rooted without relying on a distant archaeal outgroup that may introduce long-branch attraction artefacts and which some have argued may not even be an outgroup to bacteria [62, 63]. Similar considerations have motivated the use of reconciliation methods to infer the root of the archaeal [64] and eukaryotic [65] domains (Fig. 3A). The root position inferred [56] provided support for a deep divide between the Terrabacteria [66, 67] and Gracilicutes, at odds with recent inferences that the Patescibacteria/CPR are basal in the bacterial domain [68, 69]. Instead, this group of ultrasmall, reduced bacteria was recovered as sister to the Chloroflexota within the Terrabacteria, in agreement with other recent inferences [70–73]. The analysis also suggested a loss of 0.47–0.56 Mb during reductive evolution along the Patescibacteria/CPR stem lineage. Interestingly, another recent reconciliation analysis [74] indicated that reductive evolution in

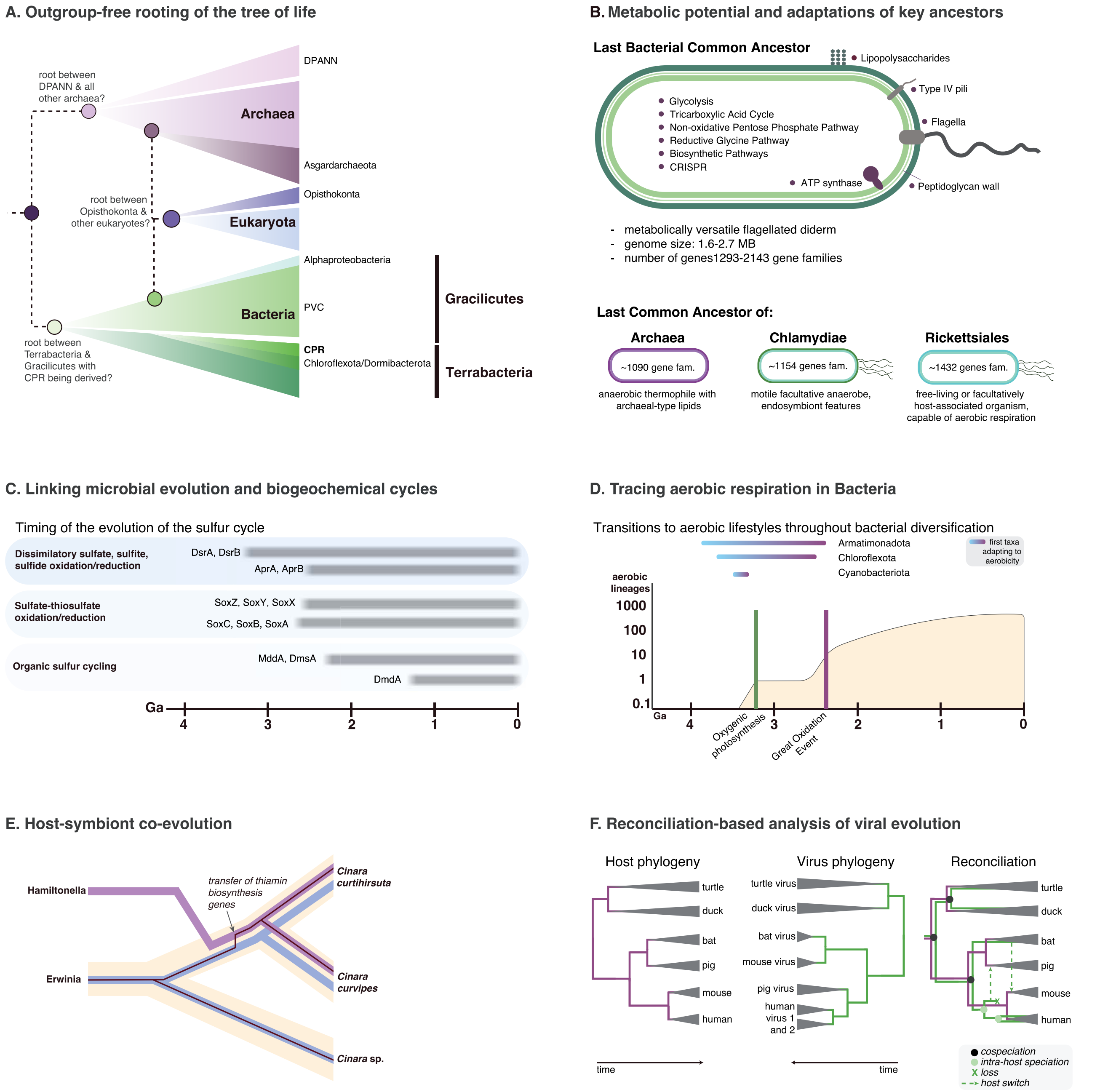
A. Outgroup-free rooting of the tree of life
root between DPANN & all other archaea?
DPANN
Archaea
Asgardarchaeota
Opisthokonta
root between Opisthokonta & other eukaryotes?
Eukaryota
Alphaproteobacteria
PVC
Bacteria
Gracilicutes
CPR
Chloroflexota/Dormibacterota
Terrabacteria
root between Terrabacteria & Gracilicutes with CPR being derived?
B. Metabolic potential and adaptations of key ancestors
Last Bacterial Common Ancestor
Lipopolysaccharides
Type IV pili
Flagella
Peptidoglycan wall
ATP synthase
Glycolysis
Tricarboxylic Acid Cycle
Non-oxidative Pentose Phosphate Pathway
Reductive Glycine Pathway
Biosynthetic Pathways
CRISPR
- metabolically versatile flagellated diderm
- genome size: 1.6-2.7 MB
- number of genes1293-2143 gene families
Last Common Ancestor of:
Archaea
~1090 gene fam.
anaerobic thermophile with archaeal-type lipids
Chlamydiae
~1154 genes fam.
motile facultative anaerobe, endosymbiont features
Rickettsiales
~1432 genes fam.
free-living or facultatively host-associated organism, capable of aerobic respiration
C. Linking microbial evolution and biogeochemical cycles
Timing of the evolution of the sulfur cycle
Dissimilatory sulfate, sulfite, sulfide oxidation/reduction
DsrA, DsrB
AprA, AprB
Sulfate-thiosulfate oxidation/reduction
SoxZ, SoxY, SoxX
SoxC, SoxB, SoxA
Organic sulfur cycling
MddA, DmsA
DmdA
Ga
4
3
2
1
0
D. Tracing aerobic respiration in Bacteria
Transitions to aerobic lifestyles throughout bacterial diversification
Armatimonadota
Chloroflexota
Cyanobacteriota
first taxa adapting to aerobicity
aerobic lineages
1000
100
10
1
0.1
Ga
4
3
2
1
0
Oxygenic photosynthesis
Great Oxidation Event
E. Host-symbiont co-evolution
Hamiltonella
transfer of thiamin biosynthesis genes
Cinara curtihirsuta
Erwinia
Cinara curvipes
Cinara sp.
F. Reconciliation-based analysis of viral evolution
Host phylogeny
turtle
duck
bat
pig
mouse
human
time
Virus phylogeny
turtle virus
duck virus
bat virus
mouse virus
pig virus
human virus 1 and 2
time
Reconciliation
turtle
duck
bat
pig
mouse
human
● cospeciation
● intra-host speciation
X loss
--> host switch

Figure 3. Inferences of microbial evolution from reconciliation analyses. (A) Phylogenetic reconciliation can be used to root trees without the use of an outgroup, which is useful when the outgroup is distant—as is the case when rooting entire domains of life—or when no outgroup is available, as is the case for the universal tree. The schematic tree shown in A is based on a recent timetree, [121] whereas ancestral gene sets were inferred in separate studies [56, 64] and are based on distinct approaches. (B) Reconciliation-based ancestral genome reconstruction has been used to infer the gene repertoires of key ancestors across the tree of life, including the archaea, bacteria, Chlamydiae [83] and Rickettsiales [82]. (C) Reconciling gene trees with a dated species tree enabled the earliest steps in the sulphur cycle to be discerned [89]. (D) Combining reconciliation analysis with machine learning classification of oxygen adaptation [95] enabled transitions in oxygen use to be mapped across the bacterial phylogeny; the earliest transitions were inferred to have occurred among the Cyanobacteriota, Chloroflexota, and Armatimonadota. (E) Three-level reconciliation models capturing host, endosymbiont, and gene trees can identify nested cases of host–symbiont coevolution, such as transfer of niche-relevant genes between endosymbionts of Cinara aphids [3, 104]. (F) Reconciliation of host and viral trees provided evidence for frequent host switching and within-host viral speciation in herpesviruses [122]; this panel is inspired by and adapted from Fig. 1 of the original study [122]. Although probabilistic reconciliation methods are powerful, results depend on the assumptions of the model used, the quality of the genomic information available and the taxon sampling, and it is likely that the view of early microbial evolution depicted here will be revised and updated as new, better-fitting models are developed and applied to these enduring questions.

Patescibacteria/CPR also continued in parallel more recently in their evolution.

The reconciliation analysis used to root Bacteria [56] indicated that roughly two-thirds of inferred gene transmissions were vertical (from ancestor to descendant), whereas one-third were horizontal on the species tree, identifying both a significant vertical and horizontal component of deep bacterial evolution [19]. Based on the reconstructed ancestral gene repertoire, the last bacterial common ancestor was rod-shaped, flagellated, motile, and diderm; i.e. it possessed a double membrane, suggesting that the monoderm (single-membrane) phenotypes of bacteria such as the Chloroflexota, Patescibacteria, Actinomycetota, and most Bacillota are the result of multiple independent losses of the outer membrane during terrabacterial evolution [71, 75].

Genome evolution associated with ecological adaptation

Reconciliation analyses have also been used to study changes in gene content during historical periods of ecological adaptation, such as changes in niche or transitions from free-living to host-associated lifestyles. One of the most striking ecological transitions in evolutionary history was the evolution of the Haloarchaea—aerobic halophiles—from anaerobic methanogenic ancestors, and several reconciliation studies have investigated the genomic changes associated with this transition [76, 77]. These analyses showed that adaptation to halophilic conditions may have occurred at least four times in different archaeal lineages, with some niche-relevant genes shared among these groups by HGT, and that a reverse transition—from salt-tolerant to more moderate conditions—may have occurred in the Hikarchaeia, close relatives of the Haloarchaea [77].

Reconciliation-based reconstruction of ancestral gene repertoires was performed in Woesearchaeota [78], a lineage within the DPANN Archaea with larger genomes and greater metabolic versatility than their sister clade, the Pacearchaeota. The authors reported that the common ancestor of the two groups also had a small genome, with the acquisition of new genes, primarily via HGT, driving an increase in metabolic versatility, and transitions between host-associated and potentially free-living lifestyles in Woesearchaeota [78]. This study demonstrates the utility of tracking genome expansion and contraction across the tree of life using reconciliation methods.

Phylogenetic reconciliation has been used [79] to trace the genomic basis of niche partitioning between non-ammonia oxidising Thaumarchaeota living in acidic topsoils and subsoils in Scotland. The divergence between topsoil and subsoil-dwelling lineages corresponded to a deep phylogenetic split. A progressive expansion of gene content was observed throughout the evolution of the topsoil lineage, resulting in the extant microbes possessing significantly larger genomes than the subsoil lineage. This work also identified the horizontal acquisition of peptidases, carbohydrate-active enzymes, and an acid-tolerant ATP synthase as factors associated with ecological adaptation to topsoil.

In a different study, the same authors [80] used this approach to infer the acquisition of metabolic traits across the ecological history of Thermoplasmatota (previously the superclass Diaforarchaea [81]), an archaeal phylum that has expanded into a myriad of diverse ecosystems, including hot spring, marine, soil, sediment, and rumen environments. The reconciliation analysis revealed that essential metabolisms such as aerobic respiration and adaptation to acidic environments were likely acquired independently multiple times in the phylum's history associated with the colonisation of new ecological niches [80].

Reconciliation methods have also been used to study the evolution of eukaryotic host association in the Chlamydiae and Rickettsiales [82, 83]. Reconstruction of the early evolution of Chlamydiae [83] suggested that the common ancestor was a facultative anaerobe that already had many of the genes needed to infect eukaryotic hosts (Fig. 3B).

The last common ancestor of Rickettsiales was predicted to be free-living or facultatively host-associated, rather than the obligate host association usually associated with members of this order. Evolution towards host association was shown to involve a general reduction in central metabolic capacity. Notably, the transition to host association corresponded with the loss of genes involved in biofilm formation and exogenous sulphate and ammonium uptake, and the gain of ATP/ADP translocase—a hallmark enzyme of energy parasitism. Furthermore, evolution towards intracellularity corresponded with a loss of amino acid biosynthesis. Conversely, evolution towards ectosymbiosis corresponded with the gain of adhesin production and export genes [82].

In addition to analyses of individual lineages, a number of reconciliation studies have focused on identifying large collections of HGT events in order to identify underlying genetic and ecological structure [84–86]. From a dataset of 960 000 trees, the parsimony method RANGER-DTL 2 was used to generate a dataset of ~2.4 million recent transfer events [86]. From these data, they observed widespread HGT with transfers inferred in 66% of gene trees. From the perspective of pangenomes [87], "accessory" genes (genes encoded by only some members of a species) were more frequently transferred than "core" genes (genes found in all members of a species). Rates of transfer also differed markedly by gene functional category and lineage, with highly abundant and co-occurring lineages exchanging the most genes [86]. These insights might inform the development of increasingly realistic reconciliation models.

Linking microbial evolution with Earth history

Present-day biogeochemical cycles are to a large extent driven by microbial metabolism, and one emerging use of phylogenetic reconciliation is to reconstruct their historical assembly by determining the origins of the associated genes and metabolic pathways. One study [88] reconciled gene trees for nitrogen-metabolising enzymes with a dated species tree using the parsimony reconciliation method AnGST [51], reconstructing the evolution of the nitrogen cycle by dating the origin of each step. Their results suggested that nitrogen fixation mediated by molybdenum-dependent nitrogenases evolved early in life's history (in the Archaean period, prior to 2.7Ga), whereas enzymes for denitrification evolved later, and proliferated widely by HGT after the Great Oxidation Event [88]. A similar analysis of key enzymes of sulphur metabolism using AnGST and another parsimony method, ecceTERA [52], indicated that energy conversion via sulphite reduction (or sulphide oxidation) was likely the earliest step in the evolution of the sulphur cycle [89]. ecceTERA [52] has also been used to map genes for phosphorus uptake and metabolism onto the tree of life, in order to reconstruct the bioavailability of phosphorus compounds through geological time [90]. The results suggested that phosphate uptake using a phosphate/sodium symporter was the earliest means of phosphorus acquisition among extant prokaryotes, consistent with the hypothesis that concentrations of environmental phosphate were relatively high in the Archaean period [91].

Reconciliation analysis was recently used to confirm that 2-methylhopane is a reliable biomarker for early cyanobacterial evolution, which had been contested based upon the presence of the biosynthetic gene in some Alphaproteobacteria [92]. A reconciliation analysis using the parsimony algorithm Notung [53] showed that the gene was acquired relatively recently by this latter group, suggesting that 2-methylhopane remains a reliable indicator of Cyanobacteria in older rocks [93].

One of the most significant biospheric transitions in Earth history was the Great Oxidation Event, when oxygen began to accumulate in the atmosphere, and the planet transitioned from a predominantly anaerobic to an oxidised world [94]. We recently combined machine learning with phylogenetic reconciliation to reconstruct the history of oxygen metabolism in Bacteria [95]. We used machine learning to learn the relationship between gene content and aerobic metabolism in modern taxa, then used reconciliation to trace the evolution of oxygen use phenotypes through time. Our analyses suggested that aerobic respiration evolved before the Great Oxidation Event but did not become widespread until the oxidation of the atmosphere [95], providing additional time constraints for inferring the geological history of Bacteria (Fig. 3D).

More broadly, several reconciliation methods have recently been developed that use the relative age constraints implied by patterns of donor-to-recipient gene transfer to improve the accuracy of molecular clock inferences for microbes, which have traditionally been hindered by the lack of an interpretable fossil record [96, 97]. Although the application of these methods is in its infancy, they promise to substantially refine the timescale for the early evolution of life.

These examples illustrate how phylogenetic reconciliation can provide new perspectives on evolutionary questions that have proven intractable based upon the fossil or biogeochemical records alone. However, the reverse is also true, where fossil or biogeochemical information can inform reconciliation—e.g. by providing age constraints that have to be taken into consideration by the reconciliation model [98], thus reflecting the interplay between microbial evolution and Earth history.

Host-symbiont coevolution

Although most attention has been focused on reconciliation of gene and species trees, phylogenetic reconciliation has the potential to be applied to other systems with similar hierarchical relationships [8]. Of potential interest to microbial ecologists is the use of phylogenetic reconciliation to model host-symbiont coevolution and the evolution of microbiome composition over time, where the symbionts are the “genes” and the host is the “species.” The approach could be used to determine the extent and timescale of host-symbiont coevolution and to identify host (or symbiont) switching events [99, 100]. This has already been demonstrated for obligate symbiont–host relationships such as *Wolbachia* and arthropod hosts [101]. In a more complex microbial ecosystem, Groussin et al. [102] used phylogenetic reconciliation to show how diet and host evolution shape gut bacteria over time, supporting a role for cospeciation in mammalian gut microbiome evolution and suggesting connections to human immune diseases. By contrast, recent reconciliation analyses by Maestri et al. [103] suggested that mammalian coronaviruses arose recently and have been horizontally transmitted among distantly related hosts, finding no evidence for long-term codiversification during mammalian evolution. A promising new approach for analysing cases of host-symbiont coevolution is the TALE algorithm [3], which implements a three-level reconciliation model capturing dependencies between host lineages, their symbiotic bacteria, and symbiont genes. Initial application of this method showed that it was better able to identify gene transfers of niche-relevant metabolic genes among bacterial symbionts of *Cinara* aphids than a simpler two-level (symbiont-gene) model [3, 104] (Fig. 3E).

Viral evolution

Reconciliation methods provide a useful framework for investigating cospeciation and coevolution between viruses and their hosts, but an interesting open question is whether, and to what extent, the conceptual model of microbial evolution implicit in reconciliation analyses (Fig. 1A) is applicable to the evolution of viruses more generally. Is it appropriate to think of viral lineages that retain coherence through time despite gene transfer or is viral evolution sufficiently dynamic that the imposition of an overarching lineage tree hinders, rather than aids, conceptual understanding, and practical phylogenetic inference? Recent advances in viral taxonomy—and the recognition of several major lineages of viruses, each united by shared core genes [105, 106]—may provide a framework with which to test these questions empirically. Recent work suggests that a lineage tree may provide a useful structure for the study of at least some lineages, including the large DNA viruses (Nucleocytoviricota, [107]). A practical challenge to reconciliation-based analyses of viral evolution is the difficulty of inferring reliable trees for viral genes, which often contain long branches that can be difficult to resolve accurately and might exhibit insufficient phylogenetic signal due to relatively short genomes in conjunction with an extremely large number of sequences [108]. However, reconciliation approaches are valuable to study virus–host coevolution and to investigate viral genome dynamics over shorter evolutionary timeframes and at shallow taxonomic levels. For example, a time-resolved reconciliation-based approach was recently implemented to determine to what extent herpesviruses evolve by coevolution (Fig. 3F), revealing that other mechanisms such as intrahost speciation, virus loss, and host switches are much more prevalent than anticipated previously (Brito et al., 2021 [122]). In another study, a phylogenetic reconciliation workflow combining the parsimony methods TreeFix-DTL [11] and RANGER-DTL [2] was used to identify recombination events in rapidly evolving viruses like SARS-CoV-2 [109].

Using machine learning to improve reconciliation analyses

Reconciliation analyses face a major scaling challenge to handle the enormous amount of genome data now available for known microbial lineages, both in the reconciliation step itself but also in the phylogenetic inference of the underlying gene trees. However, progress is rapid in this field, and the even more rapid development of machine learning methods offers new opportunities to accelerate phylogenetic and reconciliation analyses without compromising accuracy. New methods that use machine learning to rapidly select the best-fit phylogenetic model [110, 111], to optimise analysis settings [112, 113], to efficiently search tree space [114], and to inexpensively predict bootstrap support values [115] will all benefit phylogenetic inference. One reason for the efficiency of these new methods is that machine learning algorithms can predict the results of computationally expensive likelihood calculations using cheap-to-compute input features, such as parsimony (gene) trees, which exhibit high feature importance in recent studies (70%–80%; [112, 115]). An analogous approach may prove beneficial for reconciliation, where computationally cheap input features derived from parsimony reconciliations could be used to predict the results of a full probabilistic analysis.

Conclusions and prospects for progress

Phylogenetic reconciliation methods have emerged from the evolutionary biology community and are now becoming increasingly popular in the analysis of microbial genomes. As our examples demonstrate, reconciliation methods provide a natural analytical framework for studying microbial ecology and evolution that, in our view, is currently underutilized. We expect that these approaches will become more popular and deliver new insights into microbial diversity. There are underlying similarities between the substitution models used in traditional phylogenetics and probabilistic models for phylogenetic reconciliation, and just as with standard phylogenetic analyses, reconciliation studies will benefit from the development of improved models that better capture the patterns and processes of microbial evolution.

In addition to helping to speed up pipelines, it may also be possible to use machine learning in combination with reconciliation analyses to better model key aspects of microbial ecology and evolution. A current limitation of reconciliation methods (and indeed, most phylogenetic approaches) is that they do not consider coevolutionary interactions among gene families due to the computational complexity of doing so. Existing work that models gene co-occurrence and mutual avoidance patterns in modern bacterial pangenomes [116] might be extended to model interactions through time, in order to improve inference of ancestral gene complements. Machine learning may also be of use in inferring ancestral metabolic capabilities based upon reconstructed gene complements, as we demonstrated recently [95]. Beyond the origin of aerobic respiration, other metabolic transitions—such as the origin of photosynthesis—could be studied using the same approaches, raising the possibility of reconstructing ancient ecologies and biogeochemical cycles by integrating the genomic, fossil, and isotopic records, providing a powerful new statistical framework for studying microbial ecology and evolution through deep time.

Conflicts of interest

None declared.

Funding

This work was supported by the Gordon and Betty Moore Foundation through grant GBMF9741 to T.A.W., A.S., L.L.S., and G.J.S. Furthermore, G.J.S., L.L.S., and L.E. received funding from the European Union's Horizon 2020 research and innovation programme (grant agreement No. 714774, GENECLOCKS to G.J.S. and L.S. and No. 803151 to L.E.). A.S. 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), the Simons Foundation (735929LPI https://doi.org/10.46714/735929LPI)), A.A.D. and P.H. were supported by an Australian Research Council Laureate Fellowship (grant FL150100038). L.E. was also supported by a grant from the Simons Foundation (No. 812811). This work was financially supported by the European Union's Horizon Europe ERA Chair program under grant agreement No. 101087081 (Comp-Biodiv-GR). C.G.R. was supported by a Royal Society University Research Fellowship (URF150571).

Data availability

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

References

1. Lindell D, Sullivan MB, Johnson ZI *et al.* Transfer of photosynthesis genes to and from Prochlorococcus viruses. *Proc Natl Acad Sci USA* 2004;**101**:11013–8. https://doi.org/10.1073/pnas.0401526101
2. Bansal MS, Kellis M, Kordi M *et al.* RANGER-DTL 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss. *Bioinformatics* 2018;**34**:3214–6. https://doi.org/10.1093/bioinformatics/bty314
3. Menet H, Trung AN, Daubin V *et al.* Host-symbiont-gene phylogenetic reconciliation. *PeerCommunityJournal* 2023;**3**. https://doi.org/10.24072/pcjournal.273
4. Morel B, Williams TA, Stamatakis A *et al.* AleRax: a tool for gene and species tree co-estimation and reconciliation under a probabilistic model of gene duplication, transfer, and loss. *Bioinformatics* 2024;**40**:btae162. https://doi.org/10.1093/bioinformatics/btae162
5. Blanquart S, Groussin M, Le Roy A *et al.* Resurrection of ancestral malate dehydrogenases reveals the evolutionary history of Halobacterial proteins: deciphering gene trajectories and changes in biochemical properties. *Mol Biol Evol* 2021;**38**:3754–74. https://doi.org/10.1093/molbev/msab146
6. Zwaenepoel A, Van de Peer Y. Inference of ancient whole-genome duplications and the evolution of gene duplication and loss rates. *Mol Biol Evol* 2019;**36**:1384–404. https://doi.org/10.1093/molbev/msz088
7. Kapli P, Flouri T, Telford MJ. Systematic errors in phylogenetic trees. *Curr Biol* 2021;**31**:R59–64. https://doi.org/10.1016/j.cub.2020.11.043
8. Menet H, Daubin V, Tannier E. Phylogenetic reconciliation. *PLoS Comput Biol* 2022;**18**:e1010621. https://doi.org/10.1371/journal.pcbi.1010621
9. Akerborg O, Sennblad B, Arvestad L *et al.* Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc Natl Acad Sci USA* 2009;**106**:5714–9. https://doi.org/10.1073/pnas.0806251106
10. Szöllõsi GJ, Rosikiewicz W, Boussau B *et al.* Efficient exploration of the space of reconciled gene trees. *Syst Biol* 2013;**62**:901–12. https://doi.org/10.1093/sysbio/syt054
11. Bansal MS, Wu Y-C, Alm EJ *et al.* Improved gene tree error correction in the presence of horizontal gene transfer. *Bioinformatics* 2015;**31**:1211–8. https://doi.org/10.1093/bioinformatics/btu806
12. Mishra S, Smith ML, Hahn MW. reconcILS: a gene tree-species tree reconciliation algorithm that allows for incomplete lineage sorting. *bioRxiv* 2024; 2023.11.03.565544.
13. Minh BQ, Dang CC, Vinh LS *et al.* QMaker: fast and accurate method to estimate empirical models of protein evolution. *Syst Biol* 2021;**70**:1046–60. https://doi.org/10.1093/sysbio/syab010
14. Susko E, Lincker L, Roger AJ. Accelerated estimation of frequency classes in site-heterogeneous profile mixture models. *Mol Biol Evol* 2018;**35**:1266–83. https://doi.org/10.1093/molbev/msy026
15. Schrempf D, Lartillot N, Szöllösi G. Scalable empirical mixture models that account for across-site compositional heterogeneity. *Mol Biol Evol* 2020;**37**:3616–31. https://doi.org/10.1093/molbev/msaa145
16. Lake JA, Henderson E, Oakes M *et al.* Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proc Natl Acad Sci USA* 1984;**81**:3786–90. https://doi.org/10.1073/pnas.81.12.3786
17. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains archaea, bacteria, and Eucarya. *Proc Natl Acad Sci USA* 1990;**87**:4576–9. https://doi.org/10.1073/pnas.87.12.4576
18. Eme L, Spang A, Lombard J *et al.* Archaea and the origin of eukaryotes. *Nat Rev Microbiol* 2017;**15**:711–23. https://doi.org/10.1038/nrmicro.2017.133

19. Doolittle WF. Phylogenetic classification and the universal tree. *Science* 1999;**284**:2124–8. https://doi.org/10.1126/science.284.5423.2124
20. Maddison WP. Gene trees in species trees. *Syst Biol* 1997;**46**:523–36. https://doi.org/10.1093/sysbio/46.3.523
21. Diop A, Torrance EL, Stott CM *et al.* Gene flow and introgression are pervasive forces shaping the evolution of bacterial species. *Genome Biol* 2022;**23**:239. https://doi.org/10.1186/s13059-022-02809-5
22. Dagan T, Martin W. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci USA* 2007;**104**:870–5. https://doi.org/10.1073/pnas.0606318104
23. Doolittle WF. W. Ford Doolittle. *Curr Biol* 2004;**14**:R176–7. https://doi.org/10.1016/j.cub.2004.02.010
24. Lynch M, Conery JS. The origins of genome complexity. *Science* 2003;**302**:1401–4. https://doi.org/10.1126/science.1089370
25. Ohno S. *Evolution by Gene Duplication*. Berlin/Heidelberg, Germany: Springer-Verlag, 1970.
26. Gogarten JP, Kibak H, Dittrich P *et al.* Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proc Natl Acad Sci USA* 1989;**86**:6661–5. https://doi.org/10.1073/pnas.86.17.6661
27. Iwabe N, Kuma K, Hasegawa M *et al.* Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc Natl Acad Sci USA* 1989;**86**:9355–9. https://doi.org/10.1073/pnas.86.23.9355
28. Nagai K, Oubridge C, Kuglstatter A *et al.* Structure, function and evolution of the signal recognition particle. *EMBO J* 2003;**22**:3479–85. https://doi.org/10.1093/emboj/cdg337
29. Maddamsetti R, Yao Y, Wang T *et al.* Duplicated antibiotic resistance genes reveal ongoing selection and horizontal gene transfer in bacteria. *Nat Commun* 2024;**15**:1449. https://doi.org/10.1038/s41467-024-45638-9
30. Dhar R, Bergmiller T, Wagner A. Increased gene dosage plays a predominant role in the initial stages of evolution of duplicate TEM-1 beta lactamase genes. *Evolution* 2014;**68**:1775–91. https://doi.org/10.1111/evo.12373
31. Booth TJ, Bozhüyük KAJ, Liston JD *et al.* Bifurcation drives the evolution of assembly-line biosynthesis. *Nat Commun* 2022;**13**:3498. https://doi.org/10.1038/s41467-022-30950-z
32. McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* 2011;**10**:13–26. https://doi.org/10.1038/nrmicro2670
33. Rinke C, Schwientek P, Sczyrba A *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 2013;**499**:431–7. https://doi.org/10.1038/nature12352
34. Castelle CJ, Wrighton KC, Thomas BC *et al.* Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr Biol* 2015;**25**:690–701. https://doi.org/10.1016/j.cub.2015.01.014
35. Brown CT, Hug LA, Thomas BC *et al.* Unusual biology across a group comprising more than 15% of domain bacteria. *Nature* 2015;**523**:208–11. https://doi.org/10.1038/nature14486
36. Jaffe AL, Castelle CJ, Matheus Carnevali PB *et al.* The rise of diversity in metabolic platforms across the candidate phyla radiation. *BMC Biol* 2020;**18**:69. https://doi.org/10.1186/s12915-020-00804-5
37. Katinka MD, Duprat S, Cornillot E *et al.* Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* 2001;**414**:450–3. https://doi.org/10.1038/35106579
38. Albalat R, Cañestro C. Evolution by gene loss. *Nat Rev Genet* 2016;**17**:379–91. https://doi.org/10.1038/nrg.2016.39
39. Kozlov AM, Darriba D, Flouri T *et al.* RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 2019;**35**:4453–5. https://doi.org/10.1093/bioinformatics/btz305
40. Minh BQ, Schmidt HA, Chernomor O *et al.* IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* 2020;**37**:1530–4. https://doi.org/10.1093/molbev/msaa015
41. Lartillot NL, Odrigue NIR, Tubbs DAS *et al.* PhyloBayes MPI : phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst Biol* 2013;**62**:611–5. https://doi.org/10.1093/sysbio/syt022
42. Vachaspati P, Warnow T. ASTRID: accurate species TRees from internode distances. *BMC Genomics* 2015; **16** Suppl 10: S3, https://doi.org/10.1186/1471-2164-16-S10-S3.
43. Zhang C, Sayyari E, Mirarab S. ASTRAL-III: increased scalability and impacts of contracting low support branches. In: Meidanis J, Nakhleh L, (eds.), *Comparative Genomics*. Cham, Switzerland: Springer, 2017, 53–75.
44. Zhang C, Scornavacca C, Molloy EK *et al.* ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy. *Mol Biol Evol* 2020;**37**:3292–307. https://doi.org/10.1093/molbev/msaa139
45. Zhang C, Mirarab S. ASTRAL-Pro 2: ultrafast species tree reconstruction from multi-copy gene family trees. *Bioinformatics* 2022;**38**:4949–50. https://doi.org/10.1093/bioinformatics/btac620
46. Lartillot N, Brinkmann H, Philippe H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol* 2007; Suppl 1: S4, **7**, https://doi.org/10.1186/1471-2148-7-S1-S4.
47. Mirarab S, Nakhleh L, Warnow T. Multispecies coalescent: theory and applications in phylogenetics. *Annu Rev Ecol Evol Syst* 2021;**52**:247–68. https://doi.org/10.1146/annurev-ecolsys-012121-095340
48. Williams TA, Cox CJ, Foster PG *et al.* Phylogenomics provides robust support for a two-domains tree of life. *Nat Ecol Evol* 2020;**4**:138–47. https://doi.org/10.1038/s41559-019-1040-x
49. Hugenholtz P, Chuvochina M, Oren A *et al.* Prokaryotic taxonomy and nomenclature in the age of big sequence data. *ISME J* 2021;**15**:1879–92. https://doi.org/10.1038/s41396-021-00941-x
50. Chaudhary R, Bansal MS, Wehe A *et al.* iGTP: a software package for large-scale gene tree parsimony analysis. *BMC Bioinformatics* 2010;**11**:574. https://doi.org/10.1186/1471-2105-11-574
51. David LA, Alm EJ. Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* 2011;**469**:93–6. https://doi.org/10.1038/nature09649
52. Jacox E, Chauve C, Szöllősi GJ *et al.* ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics* 2016;**32**:2056–8. https://doi.org/10.1093/bioinformatics/btw105
53. Stolzer M, Lai H, Xu M *et al.* Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics* 2012;**28**:i409–15. https://doi.org/10.1093/bioinformatics/bts386
54. Sjöstrand J, Tofigh A, Daubin V *et al.* A Bayesian method for analyzing lateral gene transfer. *Syst Biol* 2014;**63**:409–20. https://doi.org/10.1093/sysbio/syu007
55. Lynch M. *The Origins of Genome Architecture*. Oxford: Oxford University Press, 2007.
56. Coleman GA, Davín AA, Mahendrarajah TA *et al.* A rooted phylogeny resolves early bacterial evolution. *Science* 2021;**372**. https://doi.org/10.1126/science.abe0511

57. Scornavacca C, Jacox E, Szöllősi GJ. Joint amalgamation of most parsimonious reconciled gene trees. *Bioinformatics* 2015;**31**: 841–8. https://doi.org/10.1093/bioinformatics/btu728
58. Boussau B, Szöllosi GJ, Duret L *et al*. Genome-scale coestimation of species and gene trees. *Genome Res* 2013;**23**:323–30. https://doi.org/10.1101/gr.141978.112
59. Morel B, Kozlov AM, Stamatakis A *et al*. GeneRax: a tool for species-tree-aware maximum likelihood-based gene family tree inference under gene duplication, transfer, and loss. *Mol Biol Evol* 2020;**37**:2763–74. https://doi.org/10.1093/molbev/msaa141
60. Eme L, Tamarit D, Caceres EF *et al*. Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature* 2023;**618**:992–9. https://doi.org/10.1038/s41586-023-06186-2
61. Wu F, Speth DR, Philosof A *et al*. Unique mobile elements and scalable gene flow at the prokaryote–eukaryote boundary revealed by circularized Asgard archaea genomes. *Nat Microbiol* 2022;**7**:200–12. https://doi.org/10.1038/s41564-021-01039-y
62. Lake JA, Servin JA, Herbold CW *et al*. Evidence for a new root of the tree of life. *Syst Biol* 2008;**57**:835–43. https://doi.org/10.1080/10635150802555933
63. Cavalier-Smith T, Chao EE-Y. Multidomain ribosomal protein trees and the planctobacterial origin of neomura (eukaryotes, archaebacteria). *Protoplasma* 2020;**257**:621–753. https://doi.org/10.1007/s00709-019-01442-7
64. Williams TA, Szöllősi GJ, Spang A *et al*. Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc Natl Acad Sci USA* 2017;**114**:E4602–11. https://doi.org/10.1073/pnas.1618463114
65. Cerón-Romero MA, Fonseca MM, de Oliveira ML *et al*. Phylogenomic analyses of 2,786 genes in 158 lineages support a root of the eukaryotic tree of life between Opisthokonts and all other lineages. *Genome Biol Evol* 2022;**14**. https://doi.org/10.1093/gbe/evac119
66. Battistuzzi FU, Feijao A, Hedges SB. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evol Biol* 2004;**4**:44. https://doi.org/10.1186/1471-2148-4-44
67. Battistuzzi FU, Hedges SB. A major clade of prokaryotes with ancient adaptations to life on land. *Mol Biol Evol* 2009;**26**:335–43. https://doi.org/10.1093/molbev/msn247
68. Hug LA, Baker BJ, Anantharaman K *et al*. A new view of the tree of life. *Nat Microbiol* 2016;**1**:16048. https://doi.org/10.1038/nmicrobiol.2016.48
69. Zhu Q, Mai U, Pfeiffer W *et al*. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains bacteria and archaea. *Nat Commun* 2019;**10**:5477. https://doi.org/10.1038/s41467-019-13443-4
70. 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* 2015;**112**:6670–5. https://doi.org/10.1073/pnas.1420858112
71. Taib N, Megrian D, Witwinowski J *et al*. Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. *Nature Ecology & Evolution* 2020;**4**:1661–72. https://doi.org/10.1038/s41559-020-01299-7
72. Martinez-Gutierrez CA, Aylward FO. Phylogenetic signal, congruence, and uncertainty across bacteria and archaea. *Mol Biol Evol* 2021;**38**:5514–27. https://doi.org/10.1093/molbev/msab254
73. Aouad M, Flandrois J-P, Jauffrit F *et al*. A divide-and-conquer phylogenomic approach based on character supermatrices resolves early steps in the evolution of the archaea. *BMC Ecology and Evolution* 2022;**22**:1. https://doi.org/10.1186/s12862-021-01952-0
74. Jaffe AL, Thomas AD, He C *et al*. Patterns of gene content and co-occurrence constrain the evolutionary path toward animal association in Candidate Phyla Radiation Bacteria. *MBio* 2021;**12**:e0052121. https://doi.org/10.1128/mBio.00521-21
75. Witwinowski J, Sartori-Rupp A, Taib N *et al*. An ancient divide in outer membrane tethering systems in bacteria suggests a mechanism for the diderm-to-monoderm transition. *Nat Microbiol* 2022;**7**:411–22. https://doi.org/10.1038/s41564-022-01066-3
76. Martijn J, Schön ME, Lind AE *et al*. Hikarchaeia demonstrate an intermediate stage in the methanogen-to-halophile transition. *Nat Commun* 2020;**11**:5490. https://doi.org/10.1038/s41467-020-19200-2
77. Baker BA, Gutiérrez-Preciado A, Rodríguez Del Río Á *et al*. Expanded phylogeny of extremely halophilic archaea shows multiple independent adaptations to hypersaline environments. *Nat Microbiol* 2024;**9**:964–75. https://doi.org/10.1038/s41564-024-01647-4
78. Huang W-C, Liu Y, Zhang X *et al*. Comparative genomic analysis reveals metabolic flexibility of Woesearchaeota. *Nat Commun* 2021;**12**:5281. https://doi.org/10.1038/s41467-021-25565-9
79. Sheridan PO, Meng Y, Williams TA *et al*. Genomics of soil depth niche partitioning in the Thaumarchaeota family Gagatemarchaeaceae. *Nat Commun* 2023;**14**:7305. https://doi.org/10.1038/s41467-023-43196-0
80. Sheridan PO, Meng Y, Williams TA *et al*. Recovery of Lutacidiplasmatales archaeal order genomes suggests convergent evolution in Thermoplasmatota. *Nat Commun* 2022;**13**:4110. https://doi.org/10.1038/s41467-022-31847-7
81. Petitjean C, Deschamps P, Lopez-Garcia P *et al*. Extending the conserved phylogenetic Core of archaea disentangles the evolution of the third domain of life. *Mol Biol Evol* 2015;**32**:1242–54. https://doi.org/10.1093/molbev/msv015
82. Schön ME, Martijn J, Vosseberg J *et al*. The evolutionary origin of host association in the Rickettsiales. *Nat Microbiol* 2022;**7**: 1189–99. https://doi.org/10.1038/s41564-022-01169-x
83. Dharamshi JE, Köstlbacher S, Schön ME *et al*. Gene gain facilitated endosymbiotic evolution of Chlamydiae. *Nat Microbiol* 2023;**8**:40–54. https://doi.org/10.1038/s41564-022-01284-9
84. Song W, Wemheuer B, Zhang S *et al*. MetaCHIP: community-level horizontal gene transfer identification through the combination of best-match and phylogenetic approaches. *Microbiome* 2019;**7**:36. https://doi.org/10.1186/s40168-019-0649-y
85. Choi Y, Ahn S, Park M *et al*. HGTree v2.0: a comprehensive database update for horizontal gene transfer (HGT) events detected by the tree-reconciliation method. *Nucleic Acids Res* 2023;**51**:D1010–8. https://doi.org/10.1093/nar/gkac929
86. Dmitrijeva M, Tackmann J, Matias Rodrigues JF *et al*. A global survey of prokaryotic genomes reveals the eco-evolutionary pressures driving horizontal gene transfer. *Nat Ecol Evol* 2024;**8**: 986–98. https://doi.org/10.1038/s41559-024-02357-0
87. Brockhurst MA, Harrison E, Hall JPJ *et al*. The ecology and evolution of Pangenomes. *Curr Biol* 2019;**29**:R1094–103. https://doi.org/10.1016/j.cub.2019.08.012
88. Parsons C, Stüeken EE, Rosen CJ *et al*. Radiation of nitrogen-metabolizing enzymes across the tree of life tracks environmental transitions in Earth history. *Geobiology* 2021;**19**:18–34. https://doi.org/10.1111/gbi.12419
89. Mateos K, Chappell G, Klos A *et al*. The evolution and spread of sulfur cycling enzymes reflect the redox state of the early Earth. *Sci Adv* 2023;**9**:eade4847. https://doi.org/10.1126/sciadv.ade4847
90. Boden JS, Zhong J, Anderson RE *et al*. Timing the evolution of phosphorus-cycling enzymes through geological time

using phylogenomics. *Nat Commun* 2024;**15**:3703. https://doi.org/10.1038/s41467-024-47914-0

91. Brady MP, Tostevin R, Tosca NJ. Marine phosphate availability and the chemical origins of life on Earth. *Nat Commun* 2022;**13**:5162. https://doi.org/10.1038/s41467-022-32815-x
92. Ricci JN, Michel AJ, Newman DK. Phylogenetic analysis of HpnP reveals the origin of 2-methylhopanoid production in Alphaproteobacteria. *Geobiology* 2015;**13**:267–77. https://doi.org/10.1111/gbi.12129
93. Hoshino Y, Nettersheim BJ, Gold DA *et al.* Genetics re-establish the utility of 2-methylhopanes as cyanobacterial biomarkers before 750 million years ago. *Nature Ecology & Evolution* 2023;**7**: 2045–54. https://doi.org/10.1038/s41559-023-02223-5
94. Lyons TW, Reinhard CT, Planavsky NJ. The rise of oxygen in Earth's early ocean and atmosphere. *Nature* 2014;**506**:307–15. https://doi.org/10.1038/nature13068
95. Davín AA, Woodcroft BJ, Soo RM, Morel B, Murali R, Schrempf D, *et al.* An evolutionary timescale for bacteria calibrated using the great oxidation event. *bioRxiv*. 2023. 2023.08.08. 552427
96. Davín AA, Tannier E, Williams TA *et al.* Gene transfers can date the tree of life. *Nature Ecology & Evolution* 2018;**2**:904–9. https://doi.org/10.1038/s41559-018-0525-3
97. Mondal A, Rangel LT, Payette JG *et al.* DaTeR: error-correcting phylogenetic chronograms using relative time constraints. *Bioinformatics* 2023;**39**. https://doi.org/10.1093/bioinformatics/btad084
98. Szöllosi GJ, Boussau B, Abby SS *et al.* Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc Natl Acad Sci USA* 2012;**109**:17513–8. https://doi.org/10.1073/pnas.1202997109
99. Groussin M, Mazel F, Alm EJ. Co-evolution and co-speciation of host-gut bacteria systems. *Cell Host Microbe* 2020;**28**:12–22. https://doi.org/10.1016/j.chom.2020.06.013
100. Perez-Lamarque B, Morlon H. Comparing different computational approaches for detecting long-term vertical transmission in host-associated microbiota. *Mol Ecol* 2023;**32**:6671–85. https://doi.org/10.1111/mec.16681
101. Bailly-Bechet M, Martins-Simões P, Szöllosi GJ *et al.* How long does Wolbachia remain on board? *Mol Biol Evol* 2017;**34**:1183–93. https://doi.org/10.1093/molbev/msx073
102. Groussin M, Mazel F, Sanders JG *et al.* Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. *Nat Commun* 2017;**8**:14319. https://doi.org/10.1038/ncomms14319
103. Maestri R, Perez-Lamarque B, Zhukova A *et al.* Recent evolutionary origin and localized diversity hotspots of mammalian coronaviruses. *elife* 2024. https://doi.org/10.7554/eLife.91745.1
104. Manzano-Maŕın A, Coeur, d'acier A, Clamens A-L *et al.* Serial horizontal transfer of vitamin-biosynthetic genes enables the establishment of new nutritional symbionts in aphids' di-symbiotic systems. *ISME J* 2020;**14**:259–73. https://doi.org/10.1038/s41396-019-0533-6
105. Koonin EV, Dolja VV, Mart K *et al.* Global organization and proposed Megataxonomy of the virus world. *Microbiol Mol Biol Rev* 2020;**84**. https://doi.org/10.1128/MMBR.00061-19
106. Koonin EV, Kuhn JH, Dolja VV *et al.* Megataxonomy and global ecology of the virosphere. *ISME J* 2024;**18**. https://doi.org/10.1093/ismejo/wrad042
107. Aylward FO, Moniruzzaman M, Ha AD *et al.* A phylogenomic framework for charting the diversity and evolution of giant viruses. *PLoS Biol* 2021;**19**:e3001430. https://doi.org/10.1371/journal.pbio.3001430
108. Morel B, Barbera P, Czech L *et al.* Phylogenetic analysis of SARS-CoV-2 data is difficult. *Mol Biol Evol* 2021;**38**:1777–91. https://doi.org/10.1093/molbev/msaa314
109. Zaman S, Sledzieski S, Berger B *et al.* virDTL: viral recombination analysis through phylogenetic reconciliation and its application to Sarbecoviruses and SARS-CoV-2. *J Comput Biol* 2023;**30**:3–20. https://doi.org/10.1089/cmb.2021.0507
110. Abadi S, Avram O, Rosset S *et al.* ModelTeller: model selection for optimal phylogenetic reconstruction using machine learning. *Mol Biol Evol* 2020;**37**:3338–52. https://doi.org/10.1093/molbev/msaa154
111. Burgstaller-Muehlbacher S, Crotty SM, Schmidt HA *et al.* ModelRevelator: fast phylogenetic model estimation via deep learning. *Mol Phylogenet Evol* 2023;**188**:107905. https://doi.org/10.1016/j.ympev.2023.107905
112. Haag J, Höhler D, Bettisworth B *et al.* From easy to hopeless-predicting the difficulty of phylogenetic analyses. *Mol Biol Evol* 2022;**39**. https://doi.org/10.1093/molbev/msac254
113. Togkousidis A, Kozlov OM, Haag J *et al.* Adaptive RAxML-NG: accelerating phylogenetic inference under maximum likelihood using dataset difficulty. *Mol Biol Evol* 2023;**40**. https://doi.org/10.1093/molbev/msad227
114. Azouri D, Abadi S, Mansour Y *et al.* Harnessing machine learning to guide phylogenetic-tree search algorithms. *Nat Commun* 2021;**12**:1983. https://doi.org/10.1038/s41467-021-22073-8
115. Wiegert J, Höhler D, Haag J *et al.* Predicting phylogenetic bootstrap values via machine learning. *bioRxiv* 2024; 2024.03.04.583288.
116. Beavan AJS, Domingo-Sananes MR, McInerney JO. Contingency, repeatability, and predictability in the evolution of a prokaryotic pangenome. *Proc Natl Acad Sci USA* 2024;**121**:e2304934120. https://doi.org/10.1073/pnas.2304934120
117. Shavit L, Penny D, Hendy MD *et al.* The problem of rooting rapid radiations. *Mol Biol Evol* 2007;**24**:2400–11. https://doi.org/10.1093/molbev/msm178
118. Bettisworth B, Stamatakis A. *RootDigger: A Root Placement Program for Phylogenetic Trees*. New York: Cold Spring Harbor Laboratory, 2020, 2020.02.13.935304.
119. 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* 2022;**71**: 959–72. https://doi.org/10.1093/sysbio/syab067
120. Kellner S, Spang A, Offre P *et al.* Genome size evolution in the archaea. *Emerging Topics in Life Sciences* 2018;**2**:595–605. https://doi.org/10.1042/ETLS20180021
121. Mahendrarajah TA, Moody ERR, Schrempf D *et al.* ATP synthase evolution on a cross-braced dated tree of life. *Nat Commun* 2023;**14**:7456. https://doi.org/10.1038/s41467-023-42924-w
122. Brito AF, Baele G, Nahata KD *et al.* Intrahost speciations and host switches played an important role in the evolution of herpesviruses. *Virus Evol* 2021;**7**:veab025. https://doi.org/10.1093/ve/veab025
123. Huerta-Cepas J, Forslund K, Coelho LP *et al.* Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol Biol Evol* 2017;**34**:2115–22. https://doi.org/10.1093/molbev/msx148
124. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 2002;**30**:1575–84. https://doi.org/10.1093/nar/30.7.1575
125. Derelle R, Philippe H, Colbourne JK. Broccoli: combining phylogenetic and network analyses for orthology assignment. *Mol Biol Evol* 2020;**37**:3389–96. https://doi.org/10.1093/molbev/msaa159

126. Edgar RC. Muscle5: high-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat Commun* 2022;**13**:6968. https://doi.org/10.1038/s41467-022-34630-w

127. Ronquist F, Teslenko M, van der Mark P *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* 2012;**61**:539–42. https://doi.org/10.1093/sysbio/sys029

128. Yang Z, Rannala B. Molecular phylogenetics: principles and practice. *Nat Publ Group* 2012;**13**:303–14. https://doi.org/10.1038/nrg3186

129. Simão FA, Waterhouse RM, Ioannidis P *et al.* BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;**31**:3210–2. https://doi.org/10.1093/bioinformatics/btv351

130. Chaumeil P-A, Mussig AJ, Hugenholtz P *et al.* GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* 2022;**38**:5315–6. https://doi.org/10.1093/bioinformatics/btac672

131. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 2019;**20**:238. https://doi.org/10.1186/s13059-019-1832-y