Rooting phylogenies

Dissertation
Submitted in fulfilment of the requirements for the degree
Doktor der Naturwissenschaften (Dr. rer. Nat.)
in the Faculty of Mathematics and Natural Sciences
of the Christian-Albrechts University Kiel

Submitted by
Fernando Domingues Kümmel Tria

Kiel, August 2018

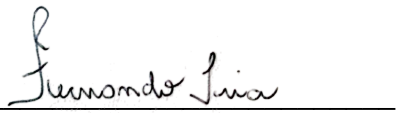
First examiner: Prof. Dr. Tal Dagan
Second examiner: Prof. Dr. Bernhard Haubold

Date of the oral examination: 12.10.2018

Declaration

I hereby declare that the thesis entitled “Rooting Phylogenies“ has been carried out in the Institute of General Microbiology at the Christian-Albrechts University of Kiel, Kiel, Germany, under the guidance of Prof. Dr. Tal Dagan and Dr. Giddy Landan. The work is original and has not been submitted in part or full by me for any degree at any other University. I further declare that the material obtained from other sources has been duly acknowledged in the thesis. My work has been produced in compliance to the principles of good scientific practice in accordance with the guidelines of the German science foundation.

Kiel, 01.08.2018



Fernando Domingues Kümmel Tria

Table of Contents

1 Abstract

Ancestor-descendent relations play a cardinal role in evolutionary theory. Those relations are determined by rooting phylogenetic trees. Existing rooting methods are hampered by evolutionary rate heterogeneity among lineages or the unavailability of auxiliary phylogenetic information. In this thesis I propose two novel rooting approaches, each approach applicable to address different research questions. In section 4, I introduce a general method to infer the roots of phylogenetic trees, without assuming prior knowledge about phylogenetic relations among the studied lineages. The method, named Minimal Ancestor Deviation (MAD), takes as input any type of unrooted tree and infers the most likely root using branch length and topological information contained in the tree. When applied to biological datasets, I show that MAD is more accurate and more robust to known confounding factors than existing methods. In the next sections, I use Ancestor Deviations (*r*) in a phylogenomic context to infer the roots of species trees, using whole genomes for the inferences. The approach is grounded in a statistical framework that evaluates all candidate roots of the underlying species tree and formally tests the relative strength of competing root hypotheses. This phylogenomic rooting approach uses information from multiple gene trees and does not require knowledge of the species tree, making it suitable for root inferences even in face of reticulated evolution.

When applied to biological datasets, our approaches reveal evidence for: 1) the origin of photosynthesis in the ocean; 2) the anaerobic and chemolithoautotrophic lifestyle of the last common ancestor of proteobacteria; and 3) the chimeric nature of modern archaea genomes.

2 Zusammenfasung (abstract in German)

Die Beziehungen zwischen Vorfahren und ihren Nachfahren spielen eine entscheidende Rolle in der Evolutionstheorie. Diese Beziehungen werden durch die Bestimmung der Wurzeln von phylogenetischen Stammbäumen ermittelt. Solche Wurzeln können durch verschiedene Methoden festgelegt werden, deren Anwendung jedoch durch heterogene Evolutionsraten oder fehlende phylogenetische Information sehr eingeschränkt ist.

In dieser Arbeit stelle ich zwei neue Wurzel-Methoden vor, welche auf unterschiedliche Fragestellungen anwendbar sind. Der erste Ansatz wird in Kapitel 4 dieser Arbeit vorgestellt und ist eine Methode zur generellen Bestimmung von phylogenetischen Wurzeln. Sie ist ohne Vorwissen über die zu untersuchenden Abstammungen anwendbar. Die neue Methode, genannt, *Minimal Ancestor Deviation* (kurz MAD), kann mit jeglicher Art von ungewurzeltem, phylogenetischem Baum durchgeführt werden. In der MAD-Methode wird mit Hilfe der Ast-Längen und topologischer Informationen des Stammbaumes die wahrscheinlichste Wurzel bestimmt. Ich zeige weiterhin, dass die MAD-Methode bei biologischen Daten ein genaueres Ergebnis produziert als bisherige Methoden und sich stabiler gegenüber Störfaktoren verhält. Im anschließenden Kapitel verwende ich die Vorfahren-Abweichungs-Statistik (*Ancestor Deviations*, *r*) in einem phylogenetischen Kontext um die Wurzel von Speziesbäumen anhand von kompletten Genomen zu bestimmen. Diese Methode basiert auf einem statistischen Vorgehen, bei welchem alle möglichen Wurzeln eines Stammbaumes direkt verglichen und evaluiert werden. Zur Bestimmung der Abstammungswurzel werden hier die Informationen von mehreren Genbäumen und nicht nur die einzelner Speziesbäume berücksichtigt. Dadurch ist diese Methode auch anwendbar, wenn eine netzartige Evolution vorliegt.

Mit den neuen Methoden aus dieser Arbeit zeige ich abschließend, 1) dass der Ursprung von Photosynthese in den Ozeanen liegt, 2) dass der letzte gemeinsame Vorfahr von Proteobakterien eine anaerobe und chemolithoautotrophische Lebensweise hatte und 3) dass Archaeen chimäre Genome, zusammengesetzt aus unterschiedlichen Spezies, aufweisen.

3 Introduction

Phylogenetic trees are used to describe and investigate the evolutionary relations between entities. A phylogenetic tree is an acyclic bifurcating graph whose topology is inferred from a comparison of the sampled entities. In the field of molecular evolution, phylogenetic trees are mostly reconstructed from DNA or protein sequences (Fitch and Margoliash 1967). Other types of data have also been used to reconstruct phylogenetic trees, including species phenotypic characteristics, biochemical makeup as well as language vocabularies (for a historical review see (Ragan 2009)). In most tree reconstruction methods the inferred phylogeny is unrooted, and the ancestral relations between the taxonomic units are not resolved. The determination of ancestor-descendant relations in an unrooted tree is achieved by the inference of a root node, which *a priori* can be located on any of the branches of the unrooted tree. The root represents the last common ancestor (LCA) from which all operational taxonomic units (OTUs) in the tree descended.

Several root inference methods have been described in the literature, differing in the type of data that can be analyzed, the assumptions regarding the evolutionary dynamics of the data, and their scalability or general applicability. The most commonly used method is the outgroup approach where OTUs that are assumed to have diverged earlier than the LCA are added to the tree reconstruction procedure (Kluge and Farris 1969). The branch connecting the outgroup to the OTUs of interest – termed ingroup - is assumed to harbor the root. Because the ingroup is assumed to be monophyletic in the resulting phylogeny, the choice of an outgroup requires prior knowledge about the phylogenetic relations between the outgroup and the ingroup. Thus, a wrong assumption regarding the outgroup phylogeny will inevitably lead to an erroneous rooted topology. Another approach, midpoint rooting, assumes a constant evolutionary rate (i.e., clock-like evolution) along all lineages, an assumption that in its strongest form, ultrametricity, equates branch lengths with absolute time (Farris 1972). In midpoint rooting the path length between all OTU pairs is calculated by summation of the lengths of the intervening branches, and the root is placed at the middle of the longest path. Midpoint rooting is expected to fail when the requirement for clock-like evolution is violated. Both outgroup and midpoint rooting can be applied independently of the tree reconstruction algorithm or the underlying type of data, with very little computational overhead. For molecular sequences and other character state data, two additional rooting methods include the root position as part of the probabilistic evolutionary models used to infer the tree topology, but at the cost of substantial increase in complexity. In the relaxed clock models approach, the evolutionary rate is allowed to vary among lineages, and the root position is optimized to produce an approximately equal time span between the LCA and all descendants (Lepage et al. 2007). In the non-reversible models approach the character transition probabilities are asymmetric and require a specification of the ancestor-descendant relation for each branch (Williams et al. 2015). Again, the root position is

optimized to maximize the likelihood of the data. Presently, both probabilistic approaches entail a significantly larger computational cost relative to the inference of unrooted trees by similar probabilistic methods. Given the cardinal role of ancestor-descendant relations in evolutionary theory, the absence of generally applicable and robust rooting methods is notable. This is in stark contrast to the wide range of methods available for the reconstruction of phylogenetic tree topologies.

Phylogenetic trees are commonly reconstructed using gene sequences (protein or DNA) to study the evolutionary history of individual gene families (i.e., homologous genes). Alternatively, whole genomes may be used to reconstruct a bifurcating species tree. The bifurcating species tree framework assumes that species descend from single ancestors through a branching process (i.e., divergence of lineages), giving rise to a tree-like diagram (Doolittle and Bapteste 2007). The root branch in the tree is the deepest branch and represents the first divergence event among the species. The subsequent branches represent more recent divergence events. The elucidation of species relations using species trees contributes to understand the properties of ancestral lineages and the chronological order of events that generated the extant biological diversity. Notable examples are: the evolution of multicellularity from single-celled organisms (West et al. 2015), the divergence of the last universal common ancestor (LUCA) into an archeon and a bacterium lineage, and later the origin of eukaryotes via endosymbiosis (Martin et al. 2015).

In prokaryotes, acquisition of genetic material is frequently through lateral gene transfers (LGT) (Popa et al. 2011), while eukaryotes experience sporadic endosymbiotic gene transfer (EGT) (Ku et al. 2015). Because of lateral events, species trees and gene trees may differ in the branching pattern and genes present in modern genomes may trace back to multiple ancestor genomes. Thus, to study the evolution of whole genomes realistically, accounting for the possibility of multiple ancestors is necessary. Despite the lateral evolutionary events, the reconstruction of species trees is commonly confined within the framework of divergences from single ancestors.

In a phylogenomic setting, the species tree is typically reconstructed from genes shared among all the species under study, termed here as complete gene families. Some gene families, however, are not present in all members of the species set, i.e., the partial gene families. Partial gene families result from gene loss in specific lineages or, alternatively, from later origin, after the divergence of the LCA. The copy number of genes may also vary. Gene families present in multiple copies in one or more species result from gene duplications and/or lateral gene transfer. Because the evolution of partial and multi-copy gene families differs from that of the species, they are commonly discarded in phylogenomic analysis targeting the reconstruction of the species tree. As a result, the reconstruction of species trees relies on single-copy, complete gene families alone (but see (Szöllosi et al. 2015) for an exception). The drawback is that the inference becomes limited to gene sets that do not represent the entirety of genomes (Medini et al. 2005). This issue tends to

become more acute the more diverse the species set is. In extreme cases no single-copy, complete gene family exists.

For species tree reconstruction, the single-copy, complete gene families may be used to generate a concatenated alignment from which a single phylogeny is reconstructed. Alternatively, the gene families may be aligned separately and multiple trees reconstructed. The trees are, then, combined in a consensus tree that represents the inference of the species tree (Thiergart et al. 2014).

Since the aforementioned approaches yield unrooted species trees, the inference of the root is required for ancestor-descendent relationship interpretations. When the species tree is reconstructed from concatenation of gene sequences, the root may be inferred during the tree reconstruction employing an outgroup (Kluge and Farris 1969), molecular clock model (Lepage et al. 2007) or non-reversible model (Williams et al. 2015). Alternatively, post-hoc rooting methods like midpoint (Farris 1972) may be applied on the species tree after reconstruction. Yet another possibility is to infer the root of the species tree from individual gene phylogenies, using complete single-copy gene families. The gene trees may be rooted using one of the aforementioned methods and the most frequent root branch in the tree sample is selected as the best species tree root. The advantage of the latter approach is that it does not require the species tree to be defined.

Here, we describe two rooting approaches. The first approach is a general rooting algorithm, termed MAD. MAD is able to infer the root of any phylogenetic tree with branch lengths including phylogenetic trees reconstructed from genes, species, morphological character, language vocabularies and any other entity that evolves according to a bifurcating tree. The second approach is a series of statistical tests to specifically target the inference of species trees roots. The inferences of species trees roots are based on information from multiple gene phylogenies and do not require *a priori* determination of the species tree.

4 Rooting trees with minimal ancestor deviation

(The content of this section was published in (Tria et al. 2017))

In this section we introduce a novel rooting method for phylogenetic trees - the Minimal Ancestor Deviation (MAD) method. MAD rooting operates on unrooted trees of contemporaneous OTUs, with branch lengths as produced by any tree reconstruction algorithm, based on any type of data, and is scalable for large datasets. No outgroup or other prior phylogenetic knowledge is required. While grounded in clock-like reasoning, it quantifies departures from clock-likeness rather than assuming it, making it robust to variation in evolutionary rates among lineages. We assessed the performance of MAD rooting in three biological datasets, one including species from the eukaryotic domain and two prokaryotic datasets of species from the cyanobacteria and proteobacteria phyla. We demonstrate that in the investigated cases MAD root inference is superior to those of the outgroup, midpoint, and the relaxed molecular clock rooting methods.

4.1 Results

4.1.1 Algorithm

The MAD method operates on binary unrooted trees and assumes that branch lengths are additive and that OTUs are contemporaneous. MAD estimates the root position by considering all branches as possible root positions, and evaluating the resulting ancestral relationships between nodes.

Before describing the algorithm, let us first define the main features of the problem (Figure 1). A rooted tree differs from its unrooted version by a single node, the root node, which is the LCA of all the OTUs considered, while internal nodes represent ancestors of partial sets of OTUs. In an n OTU unrooted tree, one can hypothesize the root node residing in any of the $2n-3$ branches. Once a branch is selected as harboring the root, the ancestral relationships of all nodes in the tree are determined. Note, however, that prior to rooting ancestral relations are unresolved, and that different root positions can invert the ancestral relations of specific internal nodes.

Under a strict molecular clock assumption (i.e., ultrametricity), the midpoint criterion asserts that the middle of the path between any two OTUs should coincide with their last common ancestor. In practice, strict ultrametricity seldom holds, and the midpoint deviates from the actual position of the ancestor node (Figure 1). The MAD algorithm evaluates the deviations of the midpoint criterion for all possible root positions and all $n(n-1)/2$ OTU pairs of the unrooted tree.

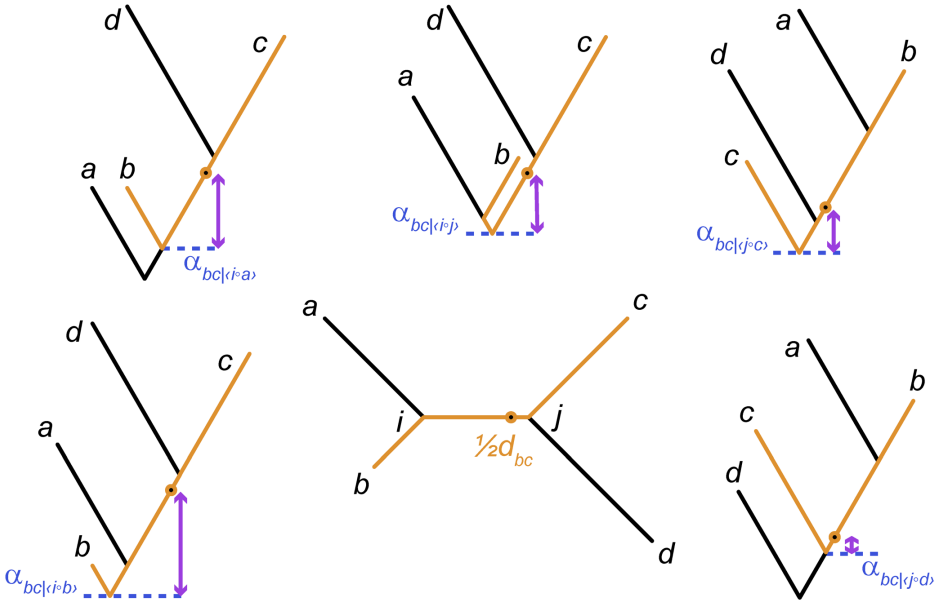


Figure 1: Schematic illustration of rooting unrooted trees. A four-OTU unrooted tree (bottom center) and the five rooted trees resulting from placing the root on each of the five branches. Yellow marks the path between OTUs *b* and *c*, and its midpoint is marked by a dot. A blue dashed line and an α mark the ancestor nodes of the OTU pair as induced by the various root positions. Purple arrows mark the deviations between the midpoint and the ancestor nodes.

Our method to estimate the root consists of: (a) considering each branch separately as a possible root position; (b) deriving the induced ancestor-descendant relationships of all the nodes in the tree; and (c) calculating the mean relative deviation from the molecular clock expectation associated with the root positioned on the branch. The branch that minimizes the relative deviations is the best candidate to harbor the root node.

Let d_{ij} be the distance between nodes i and j. For two OTUs b and c, and an ancestor node α, the distances to the ancestor are $d_{\alpha b}$ and $d_{\alpha c}$ while the midpoint criterion asserts that both should be equal to $\frac{d_{bc}}{2}$. The *pairwise relative deviation* is then defined as:

$$r_{bc,\alpha} = \left|\frac{2d_{\alpha b}}{d_{bc}} - 1\right| = \left|\frac{2d_{\alpha c}}{d_{bc}} - 1\right|,$$

(Figure 1; see Methodology in section 4.3 for the complete derivation).

For a putative root in a branch $\langle i \circ j \rangle$ connecting adjacent nodes i and j of the unrooted phylogeny, we define the *branch ancestor deviation*, $r_{\langle i\circ j\rangle}$, as the root-mean-square (RMS) of the pairwise relative deviations:

$$r_{\langle i\circ j\rangle} = \left(\overline{r_{bc,\alpha}^2}\right)^{\frac{1}{2}}$$

Branch ancestor deviations take values on the unit interval, with a zero value for exact correspondence of midpoints and ancestors for all OTU pairs, a circumstance attained only by the roots of ultrametric trees.

Branch ancestor deviations quantify the departure from strict clock-like behavior, reflecting the level of rate heterogeneity among lineages. Wrong positioning of the root will lead to erroneous identification of ancestor nodes, and apparent deviations will tend to be larger. We therefore infer the MAD root as the branch and position that minimizes the ancestor deviation $r_{\langle i \circ j \rangle}$.

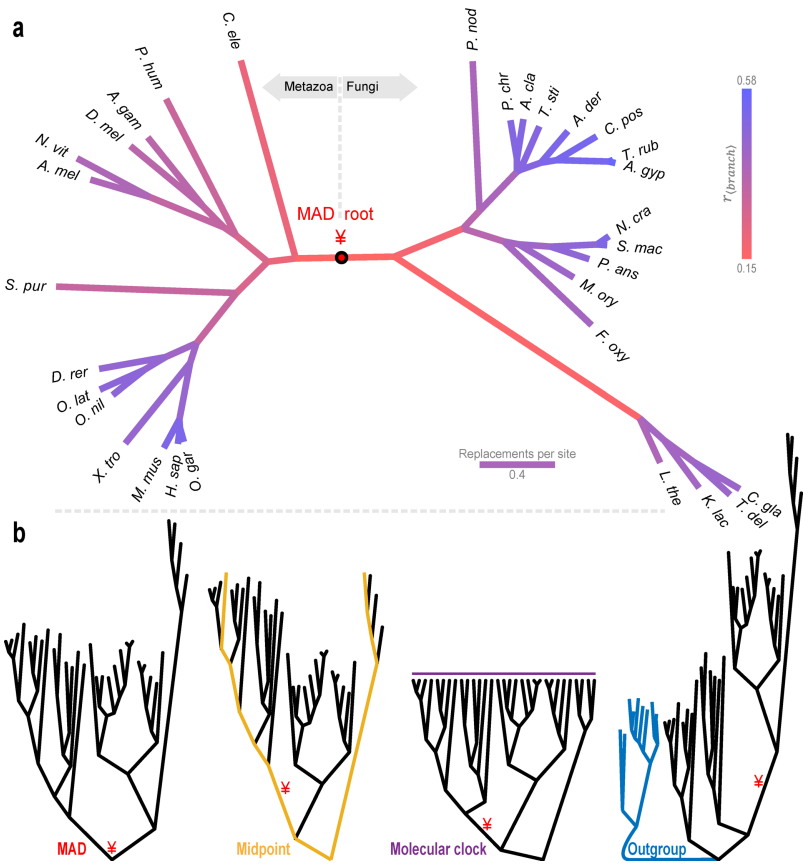


Figure 2: Minimal Ancestor Deviation (MAD) rooting illustrated with a eukaryotic protein phylogeny. **a.** An unrooted maximum-likelihood tree of trans-2-enoyl-CoA reductase protein sequences from 14 Metazoa and 17 Fungi species. Branch colors correspond to their ancestor relative deviation $r_{\langle i \circ j \rangle}$ value. The inferred root position is marked by a black circle and a red ¥ symbol. **b,** Rooted phylogenies using four alternative rooting methods, the correct root position is marked by a red ¥ symbol. The longest path of the midpoint method is marked in yellow. The molecular clock enforces ultrametricity (purple line). Ten plant outgroup OTUs are marked in blue.

We illustrate MAD rooting in Figure 2a, employing the example of an unrooted tree for 31 eukaryotic species. The minimal ancestor deviation root position is located on the branch separating fungi from metazoa. In this example, existing rooting methods place the inferred root on other branches (Figure 2b). Moreover, MAD rooting provides explicit values for all branches, thus

describing the full context of the inference. Different definitions of the deviations and averaging strategies give rise to additional MAD variants, described in 4.3 Methodology (Detailed algorithm).

4.1.2 Performance

We first consider the performance of the proposed MAD method in comparison to other rooting methods in the context of eukaryotic phylogeny. For eukaryotic sequences we expect uncertainties in root inferences to be mainly due to methodological or sampling causes rather than biological ones (e.g., reticulated evolution). We examined 1,446 trees reconstructed from protein sequences of universal orthologs in 31 opisthokonta species. The root is known to lie between fungi and metazoan (Stechmann and Cavalier-Smith 2002; Katz et al. 2012), thus giving us a clear target for the correct rooted topology. We infer root positions using the MAD method, the traditional midpoint rooting method, and the outgroup approach utilizing ten plant species as the outgroup, all based on maximum likelihood trees using PhyML (Guindon et al. 2010), as well as a Bayesian inference employing relaxed molecular clock models using MrBayes (Ronquist et al. 2012).

The four methods recover the fungi-metazoan branch as the most common inferred root position (Figure 3a; Supplementary Table 1). The MAD method identifies the correct root in 72% of the trees. The midpoint method is less consistent (61%), followed by the outgroup method (57%). The outgroup method could not be applied for 21% of the gene families, either due to the absence of plant homologs or due to multiple outgroup clusters (Supplementary Table 2). The relaxed molecular-clock method identifies the fungi-metazoa branch as the root in 36% of the trees and a neighboring branch in 34% of the trees. Neighboring branches are also found as the second most common root position in the other methods, but with much smaller frequencies (Figure 3a). The eukaryotic dataset serves as a positive control, and it demonstrates that the MAD method is accurate and consistently outperforms the existing rooting methods (see also Supplementary Tables 1 and 2 and Supplementary Figure 1).

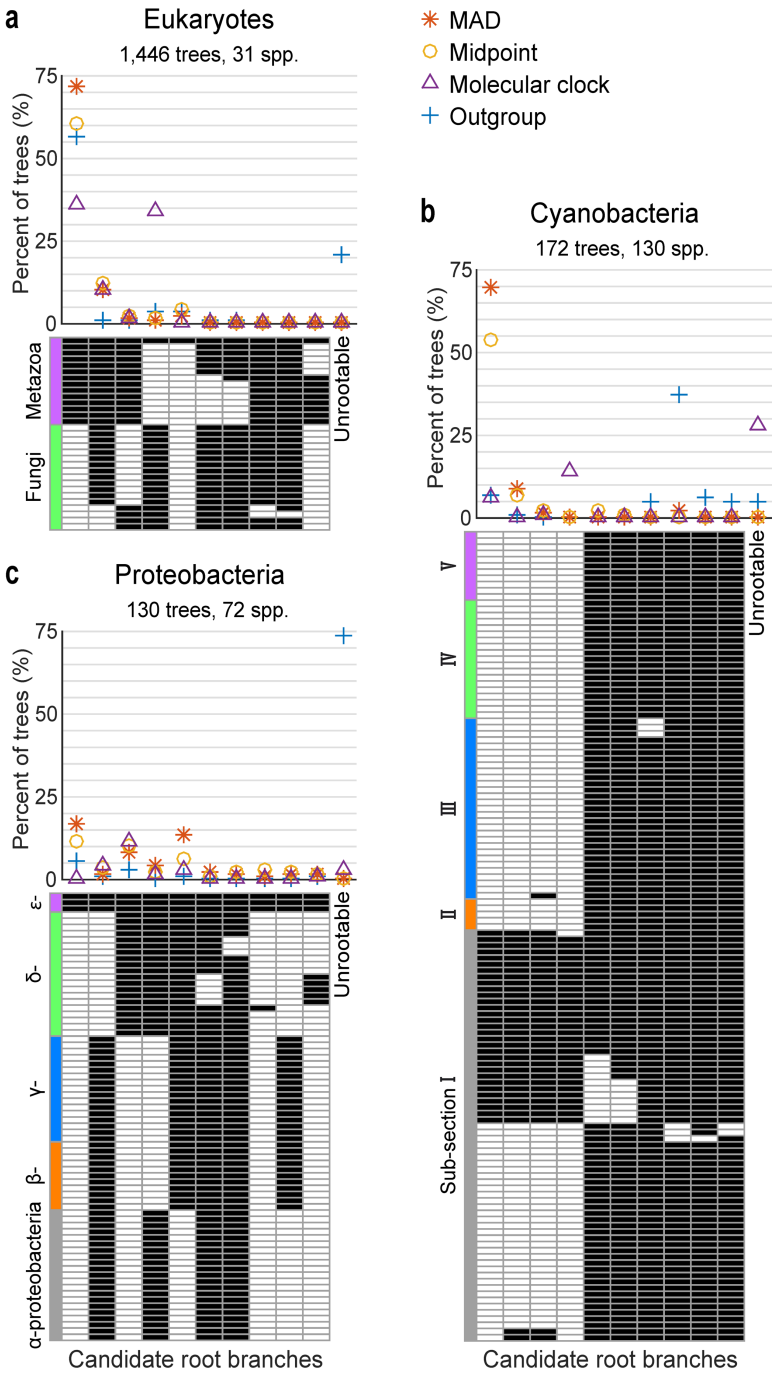


Figure 3: Root inference by four rooting methods in three datasets. Methods compared are MAD, Midpoint, Outgroup, and Molecular clock rooting. Rooting of universal protein families are summarized for **a,** Eukaryotes, **b,** Cyanobacteria and **c,** Proteobacteria (See complete list in Supplementary Table 3). (bottom) Root branches are reported as OTU splits (black and white checkered columns). The ten most frequently inferred root branches are presented (combined over the four methods). The major taxonomic groups for each dataset are indicated in color. (top) Percentage of trees with the inferred root positioned in the respective branch for each of the four methods. Rightmost position reports the proportion of unrootable trees (i.e., no outgroup orthologs, outgroup OTUs are paraphyletic, or unresolved root topology).

Rooting microbial phylogenies is more challenging because of the possibility of reticulated, non tree-like, signals (Bapteste et al. 2009). We consider the case of 130 cyanobacterial species with trees from 172 universal orthologs, using *G. violaceus* as an outgroup. *G. violaceus,* a cyanobacterium itself, is assumed to be a basal lineage (Turner et al. 1999) and serves as the traditional outgroup for other cyanobacteria (e.g. (Shih et al. 2013)). The MAD approach positions the most common root in the branch that separates a *Synechococcaceae-Prochlorococcacaea-Cyanobium* (SynProCya) clade from the remaining species, with support from 70% of the trees (Figure 3b; Supplementary Table 1). The midpoint method detects the same root position with a consistency of 54%. These values are only slightly smaller than those encountered in the eukaryotic dataset, demonstrating the robustness of MAD rooting even in the face of much deeper phylogenetic relations and possible lateral gene transfer (LGT). The second most common root position appears in just 9% of the trees, on a neighboring branch that joins two *Synochococcus elongatus* strains into the SynProCya clade. The Bayesian relaxed clock models support a neighboring branch that excludes one *Synechococcus* strain from the SynProCya clade in about 15% of the trees and produce unresolved topologies in the root position for 28% of the trees. Using *G. violaceus* as an outgroup produced a unique result by pointing to a branch separating three thermophilic *Synechococcus* strains from the rest of the phylum. This result, which is at odds with all other methods, may well stem from a wrong phylogenetic presumption of *G. violaceus* being an adequate outgroup. Using alternative outgroup species, we find variable support for the two competing root inferences, albeit always with low consistency (Supplementary Tables 1 and 2).

A more difficult rooting problem is encountered when considering highly diverse phyla. Proteobacteria groups together six taxonomic classes including species presenting diverse lifestyles and variable trophic strategies. We analyzed 130 universal gene families in 72 proteobacteria, using seven Firmicutes species as the outgroup. The MAD method produces the highest consistency, albeit at the support level of 17%, which is much lower than for the previous datasets (Figure 3c; Supplementary Table 1). The best root position is found on the branch separating epsilonproteobacteria from the remaining classes. The second most frequent branch is occurring in 14% of the trees, and the third branch in yet another 8%. All three branches occur next to each other with the second most common branch separating alphaproteobacteria from the other classes, and the third branch joining the deltaproteobacteria to the epsilonproteobacteria. These three branches are also the most frequent root braches inferred using the midpoint approach. The relaxed molecular clock approach is most frequently inferring just one of these branches as the root, the branch that separates the epsilonproteobacteria and the deltaproteobacteria from the remaining classes. We note that the outgroup approach has proved to be inapplicable for this dataset in 74% of the universal gene families.

Why does the MAD approach yield less consistent results for the proteobacteria dataset? One possibility is that this dataset presents an extreme departure from clock-likeness. We evaluate

the deviations from clock-likeness of each tree, given the inferred MAD root position, by the coefficient of variance (CV) of the distances from the root to each of the OTUs (R_{CCV}) (see Methodology). The eukaryotic dataset presents the highest level of clock-likeness, but the cyanobacterial dataset – where a consistent root branch is found – presents an even greater departure from clock-likeness than the proteobacteria dataset (Figure 4a). This shows that the lower consistency is not due to heterotachy alone and that MAD is fairly robust to departures from clock-likeness. The low support observed in proteobacteria is due to three competing branches that together account for 39% of the root inferences. This circumstance is best described as a 'root neighborhood' rather than a definite root position. To detect competing root positions for a given tree, we define the root ambiguity index, R_{AI}, as the ratio of the minimal ancestor deviation value to the second smallest value (see Methodology). This ratio will attain the value 1 for ties, i.e., two or more root positions with equal deviations, and smaller values in proportion to the relative quality of the best root position. Indeed, comparing the datasets by the distribution of the ambiguity index clearly shows that the eukaryotic dataset is the least ambiguous, while most of the trees in the proteobacteria dataset yield very high ambiguity scores (Figure 4b).

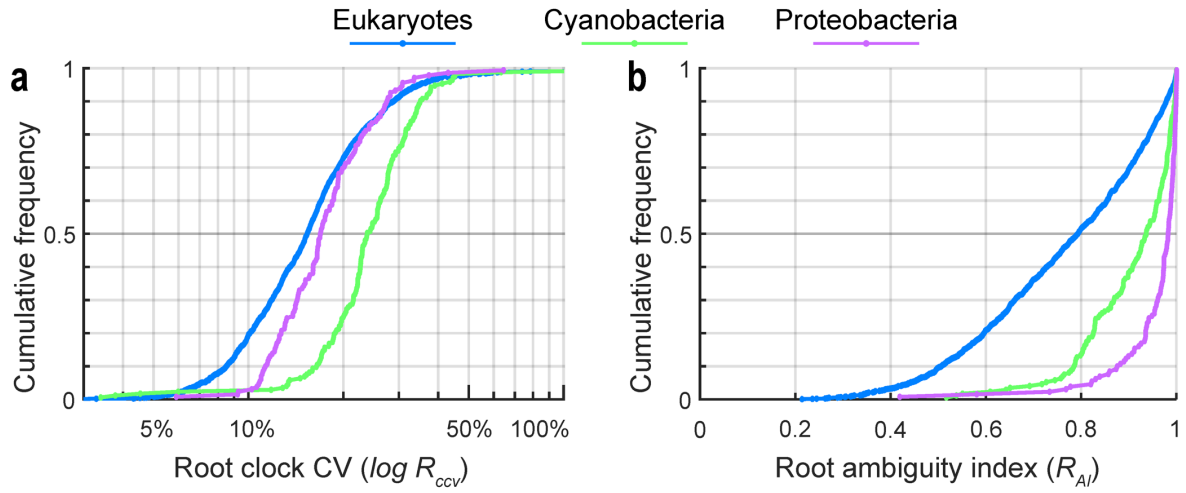


Figure 4: MAD root clock-likeness and ambiguity statistics in the three datasets.
a, Comparison of R_{CCV} distributions, which quantifies the deviation from clock-likeness, or heterotachy, associated with MAD root positions in individual trees. **b**, Comparison of the ambiguity index R_{AI} distributions for MAD root inferences.

The ambiguity observed can originate from several factors. One source of ambiguity can be due to very close candidate root positions in the tree. This situation would become more acute when the root branch is short and root positions on neighboring branches can yield comparable ancestor deviation values. Indeed, we find a significant negative correlation between the ambiguity index and the length of the root branch (normalized by tree size, Spearman ρ=-0.53; P=1.0x10^{-10}). In other words, short root branches are harder to detect.

4.2 Conclusions

Our results demonstrate that MAD rooting can outperform previously described rooting methods. Moreover, MAD operates on bifurcating trees with branch lengths, thus it is not dependent upon the type of data underlying the analysis, neither upon the tree reconstruction method or evolutionary models. MAD is also scalable; the running time of MAD is comparable to distance based tree reconstruction methods. Lastly, MAD does not depend on prior phylogenetic knowledge of outgroup species or on the availability of outgroup orthologous sequences.

The inferred MAD root for the cyanobacteria phylum implies that the last common ancestor of cyanobacteria was a unicellular organism inhabiting a marine environment. This suggests that the basic photosynthesis machinery originated in a marine environment, which contrasts with our earlier conclusions that were based on *Gloeaobacter* sp. as outgroup (Dagan et al. 2013). Alternative outgroups reproduce the MAD rooting, albeit with a lesser support. The cyanobacteria dataset shows the MAD approach to be robust to phylogenetic inference errors and possible LGT.

We introduce the concept of 'Root neighborhood' to enable the interpretation of ancestral relations in trees even in the absence of an unambiguous root position. A root neighborhood can be observed in the proteobacterial dataset, where all highly supported root positions maintain the monophyly of proteobacteria classes. The quantification of ambiguity in root inference is made possible by the evaluation of every branch as a possible root and the comparable magnitude of the ancestor deviation statistic. Thus, the MAD approach supplies a set of statistics that are intrinsically normalized, and are directly comparable between different trees. This opens the way for phylogenomic level application, with implications for the resolution of long standing species-tree conundrums. We note, however, that MAD can infer roots in any type of tree, including trees that differ from the species tree (due to paralogy or LGT, for example).

Midpoint rooting is the ultimate ancestor of the MAD approach. Three elements are new to the MAD formulation: First, the various topological pairings of midpoints to ancestor nodes; second, the exhaustive utilization of metric information from all OTU pairs (instead of just the longest path) and all possible root positions; and finally, heterotachy is embraced and explicitly quantified. Rate heterogeneity among lineages is a real phenomenon stemming from variability of the determinants of evolutionary rates: mutation rates, population dynamics and selective regimes. Thus, it is unrealistic to either assume a molecular clock or to force one by constraining the evolutionary model. The actual levels of heterotachy may appear to be even larger when a wrong position of the root is hypothesized. It is these spurious deviations that are minimized by the MAD method to infer the root position. Withstanding heterotachy is further assisted by the consideration of all OTU pairs and root positions, because lineages with exceptional rates contribute large deviations uniformly to all possible root positions.

To conclude, MAD holds promise for useful application also in other fields relying on evolutionary trees, such as epidemiology and linguistics. MAD rooting provides robust estimates of ancestral relations, the bedrock of evolutionary research.

4.3 Methodology

4.3.1 Datasets preparation

Universal protein families for the eukaryotic and proteobacteria datasets were extracted from EggNOG version 4.5 (Huerta-Cepas et al. 2016). The cyanobacteria protein families were constructed from completely sequenced genomes available in RefSeq database (O'Leary et al. 2016) (ver. May 2016), except the Melainabacteria Zag 1 genome downloaded from IMG (Markowitz et al. 2014). Species in the three datasets were selected from the available genomes so that the number of represented taxa will be as large as possible and genus-level redundancy will be reduced. The datasets are: Eukaryotes (31 opisthokonta with 10 outgroup plant species), Proteobacteria (72 species with 7 outgroup Firmicutes species), and Cyanobateria (130 species with 6 outgroup bacterial species) (See Supplementary Table 3 for the complete list of species). Outgroup species were selected according to the accepted taxonomic knowledge. EggNOG clusters with complete ingroup species-set representation were extracted, resulting in 1446 eukaryotic protein families and 130 proteobacterial protein families. For the construction of cyanobacteria protein families, at the first stage, all protein sequences annotated in the genomes were blasted all-against-all using stand-alone BLAST (Altschul et al. 1990) ver. 2.2.26. Protein sequence pairs that were found as reciprocal best BLAST hits (rBBHs) (Tatusov et al. 1997) with a threshold of E-value $\leq 1x10^{-5}$ were further compared by global alignment using needle (Rice et al. 2000). Sequence pairs having ≥30% identical amino acids were clustered into protein families using the Markov clustering algorithm (MCL) (Enright et al. 2002) ver. 12-135 with the default parameters. Protein families with complete ingroup species-set representation were retained, resulting in 172 cyanobacterial protein families.

Because in this study we are interested in universal families of orthologs only, we sorted out the paralogs from the protein families as previously described in (Thiergart et al. 2014). Of the universal protein families, 1339 eukaryotic, 85 proteobacterial and 64 cyanobacterial contained paralogous sequences, and were condensed as follows. Sequences of the protein families were aligned using MAFFT ver. v7.027b (Katoh and Standley 2013) with L-INS-i alignment strategy, and the percent of identical amino acids between all sequence pairs was calculated. Next we clustered the sequences by amino-acid identity using the single-linkage algorithm, and the largest cluster with at most a single sequence for each species was selected as a seed. Species not represented in the

seed cluster were included by the addition of the sequence with the maximal median identity to the seed cluster.

Protein sequences of the resulting universal protein families were aligned using MAFFT ver. v7.027b with L-INS-i alignment strategy. Phylogenetic trees were reconstructed using PhyML version 20120412 (Guindon et al. 2010) with the following parameters: -b -4 -v e -m LG -c 4 -s SPR. MAD rooting and midpoint rooting were performed using in-house MatLab© scripts. Molecular clock roots were inferred from phylogenies reconstructed with MrBayes ver. 3.2.3 (Ronquist et al. 2012) with the following parameters: lset rates=invgamma ngammacat=4; prset aamodelpr=fixed(wag) brlenspr=clock:uniform clockvarpr=igr; sumt contype=allcompat. Outgroup rooting was inferred from PhyML trees reconstructed from independent MAFFT alignments that include the outgroup sequences.

Code and data availability. Implementations of MAD in python, R and Matlab as well as the datasets used in this study are available through our institutional website at: https://www.mikrobio.uni-kiel.de/de/ag-dagan/ressourcen

4.3.2 Detailed algorithm

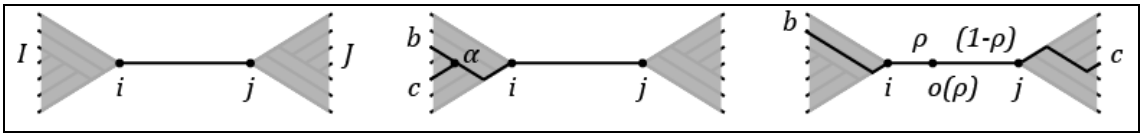
In an n OTU unrooted tree, let d_{ij} be the distance between nodes i and j, calculated as the sum of branch lengths along the path connecting nodes i and j, and thus additive by construction. For simpler exposition we will assume all branches to have a strictly positive length (i.e., $d_{ij} > 0 \; \forall \; i \neq j$). For two OTUs b and c, and a putative ancestor node α, the expected distances to the ancestor are $d_{\alpha b}$ and $d_{\alpha c}$ while the midpoint criterion asserts that both should be equal to $\frac{d_{bc}}{2}$. The resulting deviations are $\left|d_{\alpha b} - \frac{d_{bc}}{2}\right| = \left|d_{\alpha c} - \frac{d_{bc}}{2}\right|$
(see Fig. 1). To be able to summarize all OTU pairs on equal footing, we prefer to consider the deviations relative to the pairwise distance d_{bc}, and define the relative deviation as:

$$r_{bc,\alpha} = \left|\frac{2d_{\alpha b}}{d_{bc}} - 1\right| = \left|\frac{2d_{\alpha c}}{d_{bc}} - 1\right|, \qquad (1).$$

which take values on the unit interval, regardless of the magnitude of d_{bc}.

In order to compare ancestor nodes to midpoints for all pairs of OTUs, we first need to identify the last common ancestor of each OTU pair as induced by a candidate root branch. For a branch $\langle i \circ j \rangle$ connecting adjacent nodes i and j (see scheme below), we define the OTU partition $\langle I \circ J \rangle$, as:

$$I = \{terminal\ node\ k:\ \ d_{ki} < d_{kj}\},\quad J = \{terminal\ node\ k:\ \ k \notin I\}.$$



For any two OTUs lying on the same side of the putative root branch the ancestor is already present as a node in the unrooted tree, and can be identified by:

$$\alpha_{bc|\langle i\circ j\rangle} = k\colon \{d_{bc} = d_{ib} + d_{ic} - 2d_{ik}\} \quad where \quad \begin{matrix} b,c \in I; \\ k\ a\ node\ on\ the\ path\ from\ i\ to\ b \end{matrix}$$

and similarly for $b, c \in J$.

For OTU pairs straddling the candidate root branch, $b \in I, c \in J$, we first need to introduce a hypothetical ancestor node $o_{\langle i\circ j\rangle}$ with minimal deviations from the midpoints of straddling OTU pairs. Consider all possible positions $o(\rho)$ as parameterized by the relative position ρ, then $d_{io(\rho)} = \rho d_{ij}$ and $d_{jo(\rho)} = (1-\rho)d_{ij}$, and the sum of squared relative deviations is:

$$r(\rho) = \sum_{b\in I}\sum_{c\in J}\left(\frac{2d_{bo(\rho)}}{d_{bc}} - 1\right)^2 = \sum_{b\in I}\sum_{c\in J}\left(\frac{2\left(d_{bi} + \rho d_{ij}\right)}{d_{bc}} - 1\right)^2,$$

which is minimized by:

$$\rho = \sum_{b\in I}\sum_{c\in J}(d_{bc} - 2d_{bi})\, d_{bc}^{-2} \Big/ \left(2d_{ij}\sum_{b\in I}\sum_{c\in J} d_{bc}^{-2}\right) \qquad (2).$$

Since the minimizing relative position may fall outside the branch, we constrain it to the unit interval:

$$\rho_{\langle i\circ j\rangle} = min(max(0,\rho)\,,1),$$

and the position of the node $o_{\langle i\circ j\rangle}$ is given by:

$$d_{io_{\langle i\circ j\rangle}} = \rho_{\langle i\circ j\rangle} d_{ij} \quad and \quad d_{jo_{\langle i\circ j\rangle}} = (1 - \rho_{\langle i\circ j\rangle})d_{ij}.$$

The hypothetical node $o_{\langle i\circ j\rangle}$ serves as the ancestor induced by the branch for all OTU pairs straddling it: $\alpha_{bc|\langle i\circ j\rangle} = o_{\langle i\circ j\rangle}$, $b \in I, c \in J$.

For each branch we combine deviations due to all OTU pairs into the branch *ancestor deviation* score, which is defined as the root-mean-square (RMS) of the relative deviations:

$$r_{\langle i \circ j \rangle} = \left(\overline{r^2_{bc,\alpha}}\right)^{\frac{1}{2}} \quad where \quad \begin{matrix} \alpha = \alpha_{bc|\langle i \circ j \rangle}; \\ b, c \in I \cup J. \end{matrix} \qquad (3).$$

Again, $r_{\langle i \circ j \rangle}$ take values on the unit interval, with a zero value for exact correspondence of midpoints and ancestors for all OTU pairs, a condition attained only by the root nodes of ultrametric trees.

Next, we compute the ancestor deviation score for all branches. We note that the minimization equation (2), while given as an analytical point solution, can be viewed as a scan of every point in a branch. When applied to all the branches, this amounts to an exhaustive evaluation of *all points* in the unrooted phylogeny.

Finally, MAD infers the root of the tree as residing on the branch(s) with the minimal induced ancestor deviation. Let $\{\beta_1 \cdots \beta_{2n-3}\}$ be the set of branches sorted by their ancestor deviation statistic $r_{\langle \beta \rangle}$, then the root branch is β_1 and the inferred root node is:

$$^{mad}R = o_{\langle \beta_1 \rangle} \quad with\ position\ as\ defined\ in\ (2)$$

Formally, the minimal value can be attained by more than one branch, but in practice ties are very rare (not one tie in the 1748 trees analyzed here). Close competition, however, is common and can be quantified by the *root ambiguity index*:

$$R_{AI} = \frac{r_{\langle \beta 1 \rangle}}{r_{\langle \beta 2 \rangle}},$$

which take the value 1 for ties, and smaller values with increasing separation between the minimal ancestor deviation value to the second smallest value.

Since the MAD method evaluates departures from ultrametricity, it is useful to quantify the clock-likeness of the inferred root position. We define the *root clock coefficient of variation* (CV) as:

$$R_{CCV} = CV\left(d_{o_{\langle \beta_1 \rangle} b}\right) \quad with \quad b \in \{1 \cdots n\}\ OTUs \qquad (4).$$

Several elements in the preceding formulation can be modified to yield slightly different variants of MAD. We evaluated the following variants and their several combinations:

A *Definition of the pairwise deviation:*

A1 Relative deviation, equations (1) and (2) above.

A2 Absolute deviation, not normalized by the pairwise distance d_{bc}, with

$$r_{bc,\alpha} = \left| d_{\alpha b} - \frac{d_{bc}}{2} \right| = \left| d_{\alpha c} - \frac{d_{bc}}{2} \right| \quad and \quad \rho = \sum_{b \in I} \sum_{c \in J} (d_{bc} - 2d_{bi}) \,/ \left(2d_{ij} \cdot |I| \cdot |J|\right)$$

replacing equations (1) and (2).

B *Averaging of the squared pairwise deviations:*

B1 A simple mean of all $n(n-1)/2$ squared deviations, equation (3) above.

B2 Averaging occurs separately at each ancestor node for all pairs straddling it. The final score is taken as the mean of the $(n-1)$ ancestor values.

Yet other rooting variants within the conceptual framework of MAD are produced by ignoring the magnitude of deviations. In the 'Minimal Clock-CV' (MCCV) variant, hypothetical ancestor nodes $o_{\langle i \circ j \rangle}$ are retained and the resulting variation in clock-likeness, similarly to equation (4) above, is used as the branch score. Again, the branch minimizing the score is selected as the inferred root branch. In the 'Pairwise Midpoint Rooting' (PMR) variant, we omit even $o_{\langle i \circ j \rangle}$, and enumerating all pairwise paths traversing a given branch take as the score the percentage of paths with midpoints falling within the branch:

$$D_{io} = \left\{ d_{io|bc} : 0 \leq d_{io|bc} \leq d_{ij} \right\} \quad where \quad d_{io|bc} = \frac{d_{bc}}{2} - d_{ib}; \; b \in I, c \in J$$

$$PMR_{\langle i \circ j \rangle} = \frac{|D_{io}|}{|I| \cdot |J|}.$$

In this variant, the branch *maximizing* the score is the inferred root branch. Essentially, the PMR is the simplest extension of the midpoint rooting method to integrate the information from all pairwise paths.

The performances of the PMR method, the MCCV method, and of the four combinations of variants A and B are reported in Supplementary Table 1.

5 Rooting species trees

In this section I propose a phylogenomic approach to infer the roots of species trees, taking into account multiple gene trees. I present a series of statistical tests that evaluate a set of candidate roots, without assuming a specific bifurcating species tree. Each test may be applied in different analytical contexts to evaluate the relative strength of competing root hypotheses. The tests are built upon the Ancestor Deviation statistic, and extend the MAD method by the consideration of all the branches in the individual gene trees as useful information for species root inferences. The approach explores the entire information contained in whole genomes, allows for inferences of sets of likely roots (root neighborhoods) when the signal of the data is not unambiguously decisive, and does not require prior knowledge about species relations (i.e., a species tree or an outgroup). I applied this approach to four biological datasets: opisthokonta, cyanobacteria, proteobacteria and archaea. The proposed methodology retrieves the known roots for opisthokonta and cyanobacteria, while uncovering evidence for the anaerobic and chemolithotrophic lifestyle of the last common ancestor of proteobacteria. In an archaea dataset, the results reveal the chimeric nature of modern archaea genomes.

5.1 Terminology

For the sake of clarity of the exposition in this section I define below the terminology used throughout the text.

Genes

Gene family – A gene set descending from the same ancestral gene, i.e., homologous genes**.**

Complete single-copy (CSC) gene family – Gene family present as single-copy in all members of a species set.

Complete multi-copy (CMC) gene family – Gene family present in all members of a species set, but having multiple copies in at least one species.

Partial single-copy (PSC) gene family – Gene family present as single-copy in some members of the species set, but absent in others.

Partial multi-copy (PMC) gene family – Gene family present in some members of a species set, appearing in multiple copies in at least one member

Phylogenetic trees

OTUs (Operational Taxonomic Units) – The entities of study for which one wishes to reconstruct the evolutionary history. In molecular evolution, the OTUs are often genes belonging to a gene family or biological species.

Tree – A bifurcating acyclic graph representing the evolutionary history of OTUs.

Branch – An edge in a phylogenetic tree, separating the OTUs into two groups.

Split – The OTUs separation as induced by a branch in a phylogenetic tree. A split may be represented as a binary vector, indicating the OTUs grouping.

LCA – The last common ancestor of the OTUs set.

Root – The deepest internal node in a rooted phylogenetic tree, representing the LCA of all the OTUs.

Root branch – The branch in an unrooted phylogenetic tree that harbors the root node, i.e., the deepest branch in the tree.

Gene trees

Gene tree – A tree reconstructed for a gene family. In gene trees the OTUs are genes.

Species split –A branch in a gene tree that does not split the genes from the same species apart. This concept is relevant for phylogenetic trees reconstructed from multi-copy gene families, when some branches in the tree may split apart genes from the same species.

Root split – The OTUs split induced by the root branch in a gene tree.

Root ambiguity – Multiple, equally likely root branches in a gene tree.

Species trees

Species tree – A tree reconstructed for a species set. In species trees the OTUs are species.

Species partition – The division of a species set into two mutually exclusive groups. The branches in species tree are species partitions.

Root partition – The species partition as induced by the root branch in a species tree.

Root neighborhood – Multiple, equally likely root partitions for a species set. Analogous to the root ambiguity, but used in the context of species trees.

5.2 Results

5.2.1 Demonstrative datasets

I demonstrate our rooting approach using 4 biological datasets: 1) opisthokonta, 2) cyanobacteria, 3) proteobacteria, and 4) archaeabacteria. For opisthokonta and cyanobacteria the root partitions are known, whereas for proteobacteria and archaea the root partitions are still debated.

The opisthokonta dataset comprises 14 metazoa and 17 fungi species, with 117 CSC out of 18458 protein gene families. For this dataset, the known root is a partition separating fungi from metazoa species (Stechmann and Cavalier-Smith 2002; Katz et al. 2012). The cyanobacteria dataset contains 130 species, spanning five morphological sections, with 115 CSC out of 20975 protein gene families. The root partition for this dataset separates 31 unicellular species from the others (unicellular and multicellular species) (Tria et al. 2017). The ophistokonta and cyanobacteria datasets offer clear targets for root inferences, serving as positive controls, albeit with different levels of complexity. For the cyanobacteria dataset, we expect more frequent LGT and tree reconstruction artifacts than for the ophistokonta dataset.

The proteobacteria dataset comprises 72 species from 5 taxonomical classes, with 45 CSC out of 13461 protein gene families. This dataset poses a harder challenge for root inferences than cyanobacteria and ophistokonta. The results of section 4 suggested the existence of a root neighborhood of 3 branches for proteobacteria. In addition, we analyzed here a dataset of 115 archaea species from 3 phyla. The root of archaea is strongly debated, with conflicting reports in the literature (Woese et al. 1990; Waters et al. 2003; Raymann et al. 2015; Williams et al. 2017), hence it is an interesting test case for our approach. This dataset includes 26 CSC out of 9712 protein gene families.

The protein sequences of all gene families in the demonstrative datasets were aligned with MAFFT (Katoh and Standley 2013) and phylogenetic trees were reconstructed using a maximum-likelihood approach with PhyML (Guindon et al. 2010) and rooted with MAD (Details in Methodology, section 5.4).

5.2.2 Phylogenomic rooting by majority rule

Our rooting approach differs from standard ones in two aspects: 1) The consideration of gene trees reconstructed from partial and multi-copy gene families in addition to CSC gene families. 2) The evaluation of all candidate root partitions of the species set as a putative root partition. Before describing our approach, we first demonstrate the limitations of a simpler rooting approach that uses phylogenetic gene trees from CSC gene families, termed here the ***majority-rule***. Then, we show how to incorporate additional information from the CSC gene trees. The incorporation of additional

information not considered by the majority-rule enabled us to perform root inferences using statistical tests. Finally, we show how one can also consider partial and multi-copy gene families within the same statistical framework.

The majority-rule approach infers the root partition of a species set from a sample of rooted CSC gene trees. The most frequent root branch from the sample of trees (termed consensus root branch) is the inferred root partition for the species set. Note that the majority-rule approach requires that the gene trees need to be rooted prior to the analysis. In species sets with a strong root signal, the majority-rule approach is sufficient to determine a clear root partition for the species set. For example, we encountered a clear root partition with the majority-rule approach in the opisthokonta and cyanobacteria datasets, using MAD to root the individual gene trees (Figure 5). For those two datasets, the consensus root branch was inferred in more than 70% of the CSC gene trees. In contrast, the root partitions in the proteobacteria and the archaea datasets are uncertain due to the low frequency of the consensus roots as a root branch in the sample of CSC gene trees. For these two datasets, the most frequent root branch was inferred in <25% of the CSC gene trees (Figure 5). Uncertainty in inferring the root partition with the majority-rule approach arises due to competition among alternative root branches in the sample of CSC gene trees.

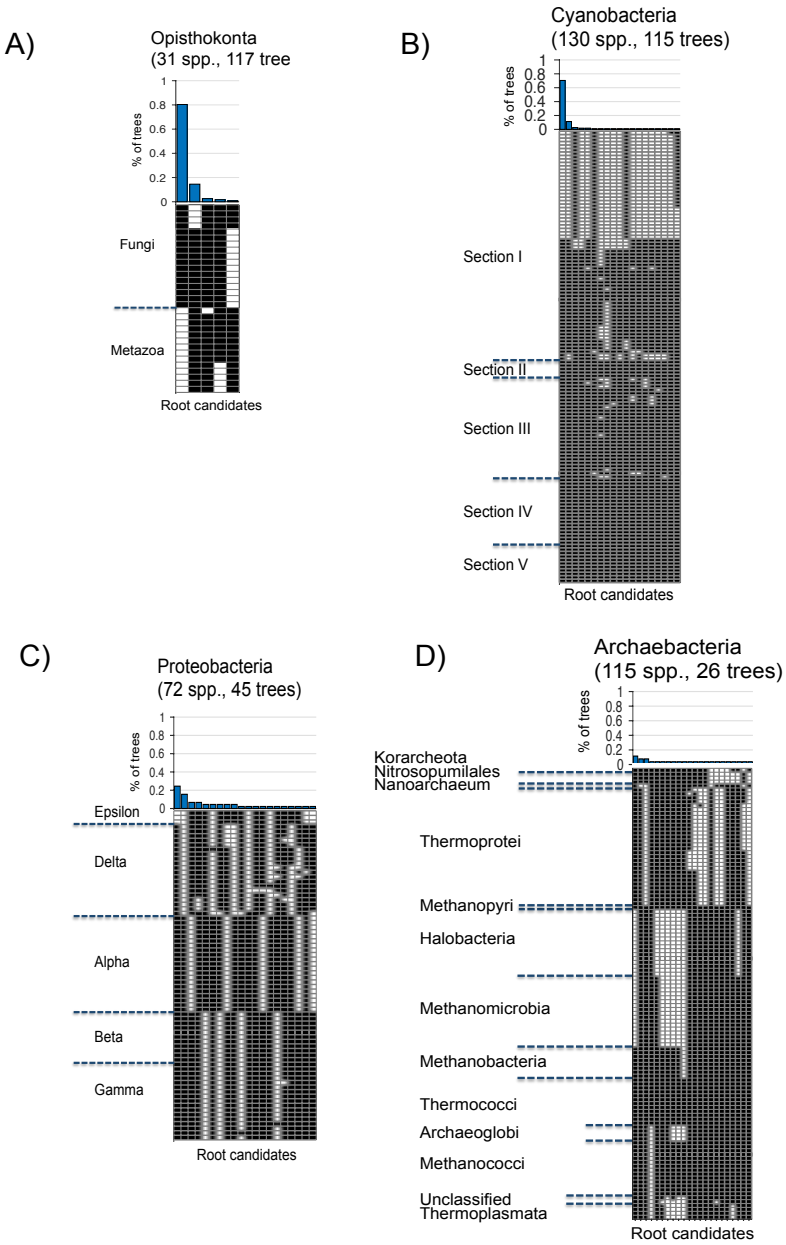


Figure 5: Consensus roots in four datasets.
MAD Rooting of CSC gene trees are summarized for A) opisthokonta, B) cyanobacteria, C) proteobacteria and D) archaea datasets**.** The inferred root branches are reported as OTU splits (white and black checkered columns) and they compose the pool of root candidates for each dataset. The percentages of trees rooted on the respective branches are displayed as bars and the taxonomic classification of the considered species is indicated on the left side. The consensus root is the branch with the highest frequency of root inferences (leftmost columns). The complete list of species composition for each dataset is given in Supplementary Table 3

5.2.3 The root support test for alternative root partitions

The majority-rule approach considers a single inferred root branch for each gene tree, and it does not account for the quality of the inferred root in the individual gene trees (e.g., ***minimal ancestor deviation*** or ***ambiguity index*** (Tria et al. 2017)).

In order to select the best root partition among all the candidates in a statistical framework, we introduce the *root support test*. The root support test considers root support values of alternative root branches in individual gene trees and, therefore, does not rely on a single root inference per gene tree. It operates by comparing the root support values for alternative root partitions, measured from a sample of gene trees. Here we calculated root support values (r) for all the branches in the sample of gene trees using MAD. Note that MAD calculates root support values in terms of ancestor deviations (r) and the smaller the r value the higher is the root support of the branch. The correspondence of branches and root partitions is immediate for CSC gene trees because all the species are represented as a single OTU. For now, we consider only CSC gene trees. Later we show how to include non-CSC in the analysis by finding the correspondence of branches and root partitions using a mapping strategy. To decide about the best root partition, we test for differences in the distribution of r values among alternative candidates and select the best-supported partition when the difference is significant.

In the simplest statistical setting, the root support test is used to compare two candidate root partitions. For the test, each considered gene tree provides measures of r values for both candidates. The candidate with significantly smaller r values in the gene tree sample is the best root partition between the two candidates. We assess the significance in the difference of paired r values using the Wilcoxon signed-rank test, considered significant when p-value<0.05. We define this test as the root pairwise test (***RPW-test***).

As an example of the RPW-test, we display the distribution of r values for two candidate root partitions from the opisthokonta dataset (Figure 6). The comparison shows that the r values for candidate 1 are significantly smaller relative to the r values for candidate 2. Thus, between the two candidates, candidate 1 is a better root partition for the underlying species set, as judged by the significance value of the RPW-test. In this example, candidate 1 corresponds to the known root partition, whereas candidate 2 corresponds to the most frequent alternative root branch in the sample of CSC gene trees.

In contrast to the opisthokonta example, we observed no significant difference in the distribution of r values for two candidate root partitions in the proteobacteria dataset. The lack of statistical difference in the distribution of r values for the two candidates indicates that the two candidate root partitions are similarly supported by the sample of gene trees.

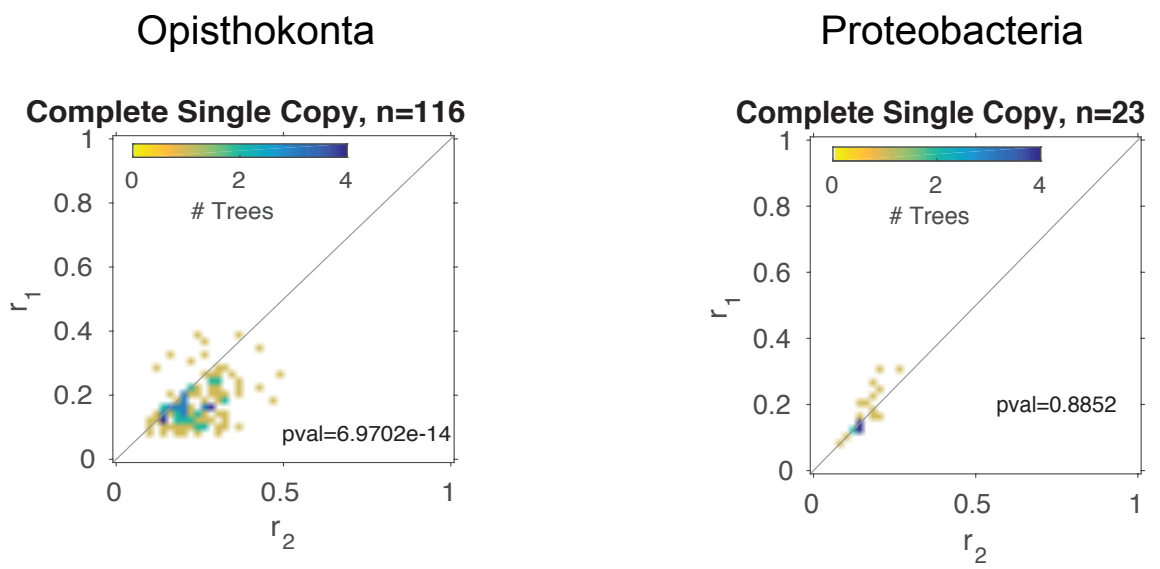


Figure 6: Ancestor deviations (*r*) calculated from CSC gene trees for two candidate root partitions from the opisthokonta and proteobacteria datasets. *r* values were calculated with MAD from a sample of CSC gene trees. r_1 denotes the *r* values for the consensus root (candidate 1) while r_2 denotes the *r* values for the most frequent alternative root branch in the CSC gene trees (candidate 2). The main diagonal is shown as a gray line and is the expected placement of gene trees with no differential support for either candidate root partition. Gene trees below the main diagonal show better support for candidate 1 (smaller *r* values correspond to better root support). The *p*-value of the RPW-test is displayed in the figure and indicates a significant better support for candidate 1 when *p*-value<0.05.

The RPW-test provides a way to test all possible pairs of root partitions for a given species set. In practice, we restricted our analysis to test only pairs of root partitions from a pool of likely candidates. For simpler exposition we assume that a pool of candidates was determined *a priori*. Later, we show how to select a pool of candidate root partitions from the sample of CSC gene trees. Given a pool of *n* candidate root partitions, all $n(n-1)/2$ candidate pairs can be tested using the RPW test. In datasets without uncertainty, the best candidate will attain significantly smaller *r* values when tested against any of the alternatives. Such is the result for the opisthokonta dataset, for which the known root partition is the best candidate among all pairwise comparisons (Table 1A). In more difficult situations the interpretation of all pairwise *p*-values is not straightforward due to the absence of a unanimous best candidate root partition. This situation is exemplified with the proteobacteria dataset where no candidate has better support than all the alternative candidates, as judged by the significance of the RPW-tests (Table 1B). The absence of a clear best candidate suggests the existence of a root neighborhood in the species set. However, it is not straightforward to determine the composition of root neighborhoods directly from the RPW-tests. Thus, a rigorous test accounting for the inference of root neighborhoods is required.

Table 1: Root pairwise tests (RPW-tests) using CSC gene trees. The *p*-values and sample sizes (in parentheses) are indicated for the differences in the distribution of *r* values between candidates (Wilcoxon signed-rank test). One sided tests with H1: values of row root candidate are smaller than values of column root candidate. The main diagonal reports the number of gene trees where the candidate is present as a branch in the CSC gene trees. The candidates were sorted according to branch frequency. In opisthokonta (A) the true root is candidate 3.

(A) Opisthokonta

	1	2	3	4	5
1	117	9.8×10^{-8} (117)	1.0 (116)	2.6×10^{-8} (115)	1.2×10^{-15} (101)
2	1.0 (117)	117	1.0 (116)	1.0 (115)	3.6×10^{-13} (101)
3	7.0×10^{-14} (116)	5.5×10^{-21} (116)	116	1.5×10^{-19} (114)	2.3×10^{-18} (100)
4	1.0 (115)	5.1×10^{-7} (115)	1.0 (114)	115	3.2×10^{-15} (100)
5	1.0 (101)	1.0 (101)	1.0 (100)	1.0 (100)	101

(B) Proteobacteria

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	45	2.9×10^{-7} (45)	0.64 (44)	1.3×10^{-5} (37)	0.53 (34)	0.89 (23)	0.021 (14)	0.94 (4)	0.81 (4)	1.0 (3)	0.88 (3)	1.0 (2)	1.0 (2)	- (1)	- (1)	- (1)	- (1)	- (1)	- (1)	- (1)
2	1.0 (45)	45	1.0 (44)	0.75 (37)	1.0 (34)	1.0 (23)	1.0 (14)	1.0 (4)	1.0 (4)	1.0 (3)	0.75 (3)	1.0 (2)	0.75 (2)	- (1)	- (1)	- (1)	- (1)	- (1)	- (1)	- (1)
3	0.36 (44)	6.4×10^{-8} (44)	44	2.5×10^{-6} (36)	0.53 (34)	1.0 (23)	1.0 (14)	1.0 (4)	0.88 (3)	0.88 (3)	0.25 (3)	1.0 (2)	1.0 (2)	- (1)	- (1)	- (1)	- (1)	- (1)	- (1)	- (1)
4	1.0 (37)	0.26 (37)	1.0 (36)	37	1.0 (29)	1.0 (17)	1.0 (12)	0.88 (3)	1.0 (4)	1.0 (3)	0.25 (3)	1.0 (2)	1.0 (2)	- (1)	- (1)	- (1)	- (0)	- (0)	- (1)	- (1)
5	0.48 (34)	2.7×10^{-7} (34)	0.48 (34)	7.6×10^{-6} (29)	34	1.0 (22)	1.0 (13)	0.75 (2)	1.0 (3)	0.50 (2)	0.25 (3)	1.0 (2)	- (1)	- (0)	- (1)	- (0)	- (0)	- (1)	- (1)	- (1)
6	0.12 (23)	1.4×10^{-5} (23)	2.5×10^{-4} (23)	1.6×10^{-4} (17)	0.0011 (22)	23	0.25 (12)	- (0)	- (1)	- (1)	- (1)	- (0)	- (0)	- (0)	- (1)	- (0)	- (0)	- (1)	- (0)	- (0)
7	0.98 (14)	6.1×10^{-5} (14)	0.0015 (14)	2.4×10^{-4} (12)	0.0034 (13)	0.78 (12)	14	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)
8	0.13 (4)	0.063 (4)	0.063 (4)	0.25 (3)	0.50 (2)	- (0)	- (0)	4	- (0)	- (0)	0.50 (2)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)
9	0.31 (4)	0.063 (4)	0.25 (3)	0.063 (4)	0.13 (3)	- (1)	- (0)	- (0)	4	- (0)	- (0)	- (1)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)
10	0.13 (3)	0.13 (3)	0.25 (3)	0.13 (3)	0.75 (2)	- (1)	- (0)	- (0)	- (0)	3	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)
11	0.25 (3)	0.38 (3)	0.88 (3)	0.88 (3)	0.88 (3)	- (1)	- (0)	0.75 (2)	- (0)	- (0)	3	- (0)	- (0)	- (0)	- (1)	- (0)	- (0)	- (0)	- (0)	- (0)
12	0.25 (2)	0.25 (2)	0.25 (2)	0.25 (2)	0.25 (2)	- (0)	- (0)	- (0)	- (1)	- (0)	- (0)	2	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)
13	0.25 (2)	0.50 (2)	0.25 (2)	0.25 (2)	- (1)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	2	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)
14	- (1)	- (1)	- (1)	- (1)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	1	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)
15	- (1)	- (1)	- (1)	- (1)	- (1)	- (1)	- (0)	- (0)	- (0)	- (0)	- (1)	- (0)	- (0)	- (0)	1	- (0)	- (0)	- (0)	- (0)	- (0)
16	- (1)	- (1)	- (1)	- (1)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	1	- (0)	- (0)	- (0)	- (0)
17	- (1)	- (1)	- (1)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	1	- (0)	- (0)	- (0)
18	- (1)	- (1)	- (1)	- (0)	- (1)	- (1)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	1	- (0)	- (0)
19	- (1)	- (1)	- (1)	- (1)	- (1)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	1	- (0)
20	- (1)	- (1)	- (1)	- (1)	- (1)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	1

An alternative statistical framework to the RPW-test is the comparison of each root partition to all the alternatives simultaneously. In this framework, each root partition is tested only once in a one-against-all manner. We term this approach as the root one-against-all test (***ROA-test***). The ROA-test consists of comparing the distribution of *r* values for one root partition to the best selection of *r* values among all the other candidates, as measured from a sample of gene trees. Each gene tree provides one *r* value for the partition under consideration and one *r* value for the best alternative root partition. Note that ROA-test permits the best alternative root partition to vary across gene trees. We test for differences in the magnitude of paired *r* values using the Wilcoxon-signed rank test. The significance of the test reflects the quality of the root partition under consideration relative to all the other root partitions. Testing each root partition only once reduces the number of total tests from $n(n-1)/2$ to n, in comparison to the RPW-tests. Additionally, the comparison of *r* values for one partition against the best selection of *r* values among the alternatives, makes the test more conservative than RPW-tests.

The ROA-test provides an alternative statistical framework to the RPW-test for selecting the best root partition among all the candidate root partitions. In species sets without uncertainty regarding the true root partition, a single candidate will result in one significant ROA-test. Such is the case for the opisthokonta and cyanobacteria datasets, where only the known root partitions result in significant ROA-tests (Table 2). In species sets with uncertainty regarding the root partition, however, no candidate will be preferred over all the others. Such is the case for the proteobacteria and archaea datasets where no candidate attains a significant ROA-test (Table 2). For species sets with uncertainty, we need further analysis for either disambiguation or inference of root neighborhoods.

Table 2: ROA-test calculated using CSC gene trees in the 4 demonstrative datasets.
The root candidate IDs follow the root partition order as displayed in Figure 5 (left to right) and the *p*-values in green shade indicate the significant results (*p*-value<0.05).

Partition ID	**Opisthokonta**	**Cyanobacteria**	**Proteobacteria**	**Archaea**
1	4.9×10^{-11}	6.1×10^{-9}	1.0	1.0
2	1.0	1.0	1.0	1.0
3	1.0	1.0	1.0	1.0
4	1.0	1.0	1.0	1.0
5	1.0	1.0	1.0	1.0
6	-	1.0	1.0	1.0
7	-	1.0	1.0	1.0
8	-	1.0	1.0	1.0
9	-	1.0	1.0	1.0
10	-	1.0	1.0	1.0
11	-	1.0	1.0	1.0
12	-	1.0	1.0	1.0
13	-	1.0	1.0	1.0
14	-	1.0	1.0	1.0
15	-	1.0	1.0	1.0
16	-	1.0	1.0	1.0
17	-	1.0	1.0	1.0
18	-	1.0	1.0	1.0
19	-	1.0	1.0	1.0
20	-	-	1.0	1.0
21	-	-	-	1.0
22	-	-	-	1.0

In order to enable disambiguation and inferences of root neighborhoods, we propose an iterative procedure that we term root partition elimination (***RPE***). Our goal with the RPE is to start with a large root neighborhood comprising all candidate root partitions, followed by iterative elimination of the worst candidate in order to improve the overall quality of the root neighborhood. For the RPE we first need to sort all the root partitions in order of root partition quality. Here, we sort the root partitions according to the significance values of the ROA-tests, thus reflecting the quality of the root partition under consideration relative to all the other candidates. In each iteration of the RPE procedure, we test the *r* values calculated for the worst root partition against the best selection of *r* values among the root partitions with better root quality, i.e., those with smaller *p*-values from the ROA-tests. We eliminate the worst root partition if the *r* values are significantly larger (worse) than the *r* values of the remaining partitions (Wilcoxon signed-rank test, FDR<0.05). We repeat the iteration until a single root partition is left or until the test is no longer significant. When a single root partition is left after the RPE procedure, then the inference is of a strict root and the candidate left is

the best root partition for the species set. When multiple root partitions are left, then the inference is of a root neighborhood, composed of all the remaining root partitions after the RPE procedure. We introduce pseudocounts for root candidates that are missing in the gene trees. The pseudocount is the maximal *r* value in the gene tree where the root candidates are absent. Table 3 demonstrates the RPE procedure for the opisthokonta and cyanobacteria datasets.

We noticed that the cyanobacteria dataset presents a challenge for the RPE procedure without the pseudocounts, since certain candidate root partitions are rarely found as a branch in the sample of CSC gene trees. Such spurious root candidates attain a small *r* sample size that leads to the lack of statistical power in specific iterations of the RPE (Table 3B). Consequently, the iterative exclusion of candidates is interrupted prematurely, rendering inferences of apparently large root neighborhoods for this dataset. The pseudocounts circumvent these biases originating from small sample sizes, as observed in the cyanobacteria dataset (Table 3B). We note that the use of pseudocounts does not change the inferred root for the opisthokonta dataset (Table 3A).

Table 3: Root inference with the RPE procedure for A) opisthokonta and B) cyanobacteria CSC gene trees.
The candidate root partitions are sorted according to the *p*-values values from the ROA-tests (second column). The FDR adjusted *p*-value of the RPE procedure is presented in the 'RPE' column and considered significant when FDR<0.05. The RPE procedure was performed with and without the inclusion of pseudocounts for missing root candidates in the gene trees (see text). The 'Sample size' column indicates the number of gene trees considered for the RPE tests. The rows in green shades display the inferred root partition.

A) Opisthokonta

		With pseudocounts		Without pseudocounts	
Partition ID	ROA tests	RPE	Sample size	RPE	Sample size
1	9.4×10^{-12}	-	-	-	-
2	1.0	4.6×10^{-13}	117	7.0×10^{-14}	116
3	1.0	3.9×10^{-21}	117	1.1×10^{-20}	117
4	1.0	3.9×10^{-21}	117	1.6×10^{-18}	101
5	1.0	3.9×10^{-21}	117	8.4×10^{-21}	115

B) Cyanobacteria

		With pseudocounts		Without pseudocounts	
Partition ID	ROA tests	RPE	Sample size	RPE	Sample size
1	6.1×10^{-9}	-	-	-	-
2	1.0	1.3×10^{-16}	113	-	-
3	1.0	8.3×10^{-19}	114	-	-
4	1.0	1.5×10^{-20}	114	-	-
5	1.0	4.5×10^{-19}	115	-	-
6	1.0	1.0×10^{-20}	115	-	-
7	1.0	7.6×10^{-21}	115	-	-
8	1.0	7.6×10^{-21}	115	-	-
9	1.0	7.6×10^{-21}	115	-	-
10	1.0	7.6×10^{-21}	115	-	-
11	1.0	7.6×10^{-21}	115	0.88	3
12	1.0	7.6×10^{-21}	115	0.0028	16
13	1.0	7.6×10^{-21}	115	9.8×10^{-4}	11
14	1.0	7.6×10^{-21}	115	7.7×10^{-6}	33
15	1.0	7.6×10^{-21}	115	5.8×10^{-7}	58
16	1.0	7.6×10^{-21}	115	3.0×10^{-8}	39
17	1.0	7.6×10^{-21}	115	1.3×10^{-10}	53
18	1.0	7.0×10^{-21}	115	4.7×10^{-18}	101
19	1.0	7.0×10^{-21}	115	1.5×10^{-20}	113

5.2.4 Phylogenetic signal from partial and multi-copy gene trees

Existing phylogenomic rooting approaches are hampered by the limited number of CSC gene families in biological datasets. In our four datasets the CSC gene families comprise less than 0.7% of the total number of gene families (Table 4). In more extreme cases, of large species sets with deep phylogenetic relations, it is even possible that none of the gene families is CSC.

Our root inference approach allows for the consideration of gene trees reconstructed from non-CSC gene families. Non-CSC gene families are those with partial species composition or present in multiple copies in one or more species. The interpretation of phylogenetic signal from non-CSC gene trees is challenging because the OTUs set differs from the species composition of the root partitions. Here we propose a mapping strategy to find branches in gene trees that correspond to root partitions.

Table 4: Distribution of gene family categories in the 4 species sets.
Percentage of each gene family category is shown. Absolute numbers are indicated in parenthesis.

	CSC	CMC	PSC	PMC	Total
Opisthokonta	0.63% (117)	7.19% (1328)	54.64% (10085)	37.53% (6928)	100% (18458)
Cyanobacteria	0.55% (115)	0.27% (57)	77.64% (16285)	21.54% (4518)	100% (20975)
Proteobacteria	0.33% (45)	0.63% (85)	63.12% (8496)	35.92% (4835)	100% (13461)
Archaea	0.27% (26)	0% (0)	50.04% (4860)	49.69% (4826)	100% (9712)

In CSC gene trees each branch in the tree correspond to a unique root partition. The branch splits of CSC gene trees are identical to their corresponding root partitions (Figure 7A).

In partial gene trees, not all the species are represented as OTUs. In order to find the branches in a partial gene tree that correspond to the root partitions, we reduce the root partitions by removing the species that are missing in the gene tree (Semple and Steel 2000). We find the correspondence between branches in partial gene trees and root partitions by mapping the OTU splits from partial gene tree to the reduced root partitions (Figure 7B). Notably, the same branch of a partial gene tree may be identical to the reduced versions of two or more root partitions. Moreover, partial gene families may be uninformative in respect to specific root partitions when all the species from one of the sides of the root partition lack the gene. Partial trees from uninformative gene families cannot have a branch split identical to the reduced version of the root partition. In our tests, we discarded partial gene trees when the same branch split mapped onto the two root partitions being tested and partial genes uninformative to either partition, since they cannot contribute to distinguish the best candidate root.

In multi-copy gene trees one or more species are represented multiple times as OTUs (Swenson and El-Mabrouk 2012). Each branch of a multi-copy gene tree splits the OTUs into two groups. The groups may be mutually exclusive or overlapping in terms of species. Mutually exclusive splits can be mapped to specific root partitions. On the other hand, overlapping splits cannot be identical to any root partition (Figure 7C).

Mapping of splits from partial multi-copy gene trees entail both operations: identification of mutually exclusive splits and reduction of root partitions.

Once we find the branches corresponding to the root partitions, we can use the *r* values measured from non-CSC gene trees in the same manner as for CSC gene trees (see previous section). We also use pseudocounts for missing candidate root partitions in non CSC gene trees. For non-CSC gene trees, pseudocounts are introduced only when the gene family is informative regarding the missing root partitions, that is, when species from both sides of the root partition contain a copy of the gene. Complete gene families are always informative to all the candidate root partitions. However, partial gene families may be informative to specific candidates.

In the next sections I present the results for root inferences using CSC and non-CSC gene trees separated and in combination, for the 4 demonstrative datasets.

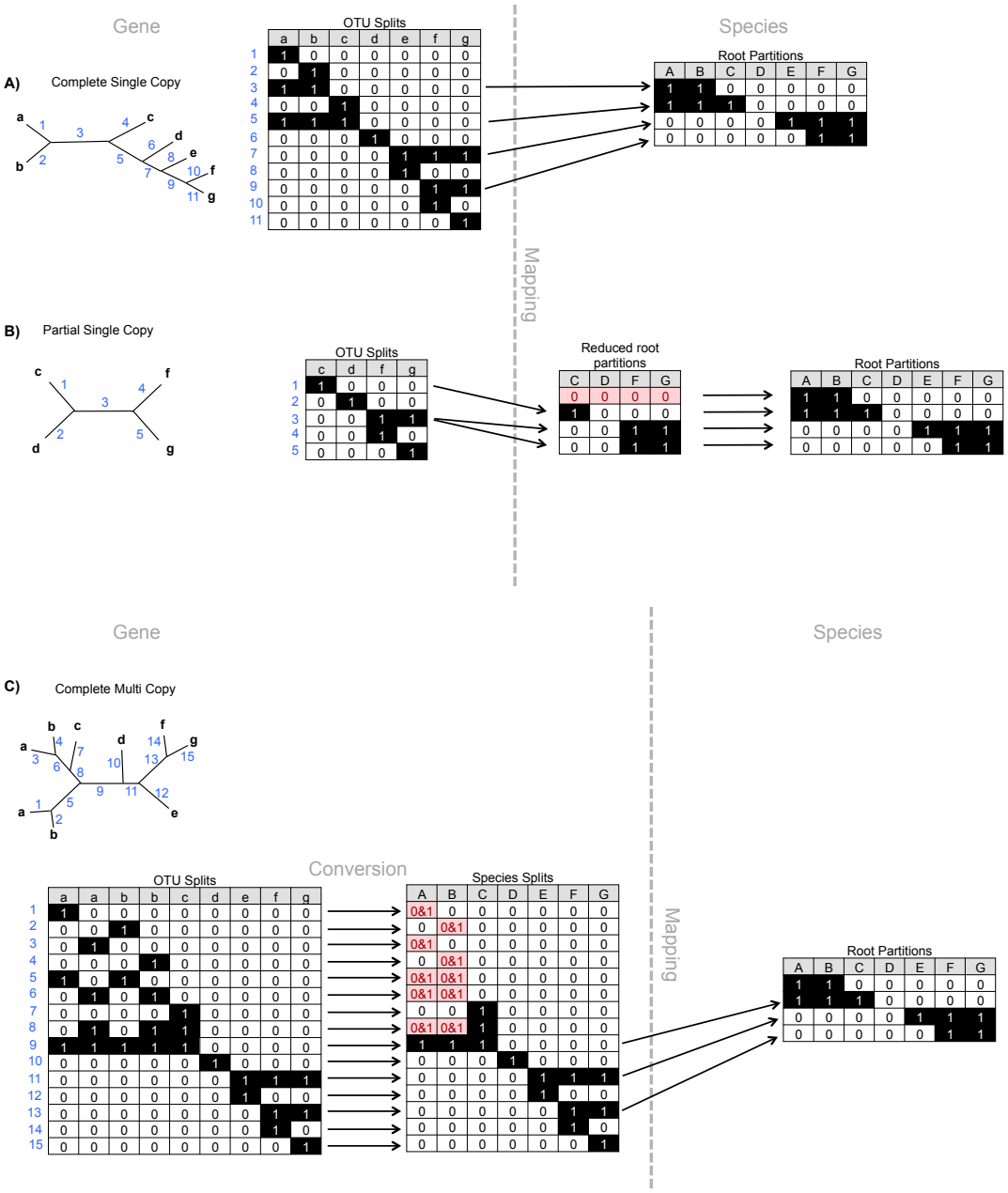


Figure 7: Mapping of gene trees splits onto species root partitions. A) Complete Single-copy, B) Partial, and C) Multi-copy. The branches in the gene trees are coded as binary splits of the OTUs. For the CSC gene tree the OTU splits can be directly mapped onto root partitions (arrows). For the partial gene tree, the root partitions are first reduced to include only species present in the gene tree. The rows shaded in red in the reduced matrix are root partitions for which the gene tree is not informative (see text). For the multi-copy gene tree the OTU split matrix is first converted into species splits. The cells shaded in red are species that appear in both sides of the OTU split, i.e., overlapping splits. Overlapping splits cannot map onto root partitions and are, hence, discarded. Mutually exclusive splits can be mapped onto the root partitions.

5.2.5 Root inferences in biological datasets

Our root inference approach provides a way to evaluate all possible root partitions of a species set. In practice, testing all possible root partitions is computationally intractable. Therefore, we restrict the analysis to a pool of likely candidates. For each dataset, we defined the pool of candidates as the set of root branches from the CSC gene trees, as inferred with MAD (see Figure 5).

Opisthokonta

The majority-rule approach applied on opisthokonta CSC gene trees reveals a clear consensus root branch. A total of 80% of CSC gene trees have the inferred MAD root on the branch corresponding to the known root partition, separating fungi from metazoa (Figure 5). The ROA-tests for this dataset show that the magnitude of *r* values calculated from CSC gene trees is significantly smaller for the known root partition in comparison to all the other candidates (Figure 8). Hence, the known root partition has a significantly better root support in the sample of CSC gene trees than all the other candidates. No other candidate root partition attained a significant *p*-value in the ROA-tests (Table 5). Notably, the trend of the ROA-tests was reproduced when we used the *r* values calculated from gene trees with paralogy and partial species sets (i.e., non-CSC gene trees). Combining the information from all gene trees increased the power of the analysis due to considerable increase in sample size (i.e., gene trees), as judged by the significance of the ROA-tests (Table 5, Figure 8).

A single significant root partition in the ROA-tests provides clear evidence for the best root partition in the dataset. The best candidate being the one with significant ROA-test. Such situations require no further analysis such as the RPE procedure. The RPE was designed to deal with uncertain datasets, when no candidate root partition attains a significant result in the ROA-tests. Datasets with clear root partitions, however, offer an opportunity to validate the accuracy of the RPE. For the opisthokonta dataset, the RPE inferred the known root partition as a single root, across all the samples of gene trees, i.e., CSC, CMC, PSC, PMC, and all trees combined (Table 6). It is noteworthy that for the each RPE iteration we consider only gene trees that contained branches corresponding to at least one of the candidates left in the pool. Therefore, the number of evaluated trees may differ between iterations of the RPE (Table 6). Taken together, these results supply a validation for the accuracy of our root inference approach.

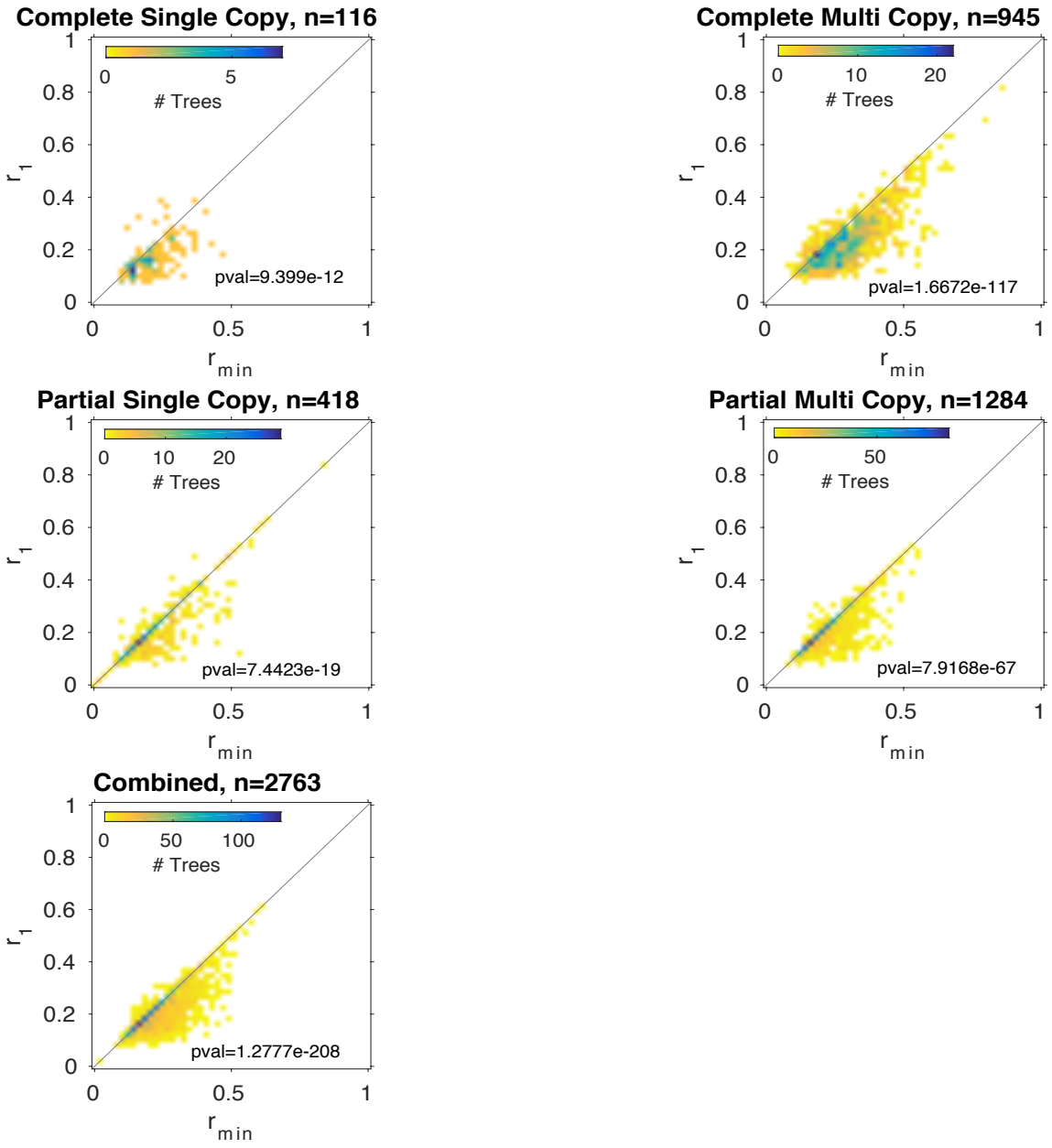


Figure 8: ROA-test for the consensus root branch in the opisthokonta dataset, across all gene family categories. The r_1 are ancestor deviation values corresponding to the consensus root branch, as measured from the sample of gene trees. Panes report different gene tree samples and *n* refers to the total number of gene trees included in the test. The significance in the difference of *r* values distribution between the consensus root (r_1) and all other candidates (r_{min}) was assessed with the Wilcoxon signed-rank test (*p*-value is indicated). The main diagonal is shown as a gray line and is the expected placement of gene trees with no differential support for the consensus root branch. Gene trees below the main diagonal show preferential support to the consensus root.

Table 5: ROA-test for all candidate root partitions in the opisthokonta dataset. The tests were performed across all samples of gene trees. The nominal *p*-values of the tests are indicated and the cells shaded in green show significant tests (*p*-value<0.05).

	With pseudocounts					Without pseudocounts				
	CSC	CMC	PSC	PMC	Combined	CSC	CMC	PSC	PMC	Combined
1	4.9×10^{-11}	1.6×10^{-12}	1.3×10^{-15}	0.0015	9.4×10^{-25}	9.4×10^{-12}	1.7×10^{-117}	7.4×10^{-19}	7.9×10^{-67}	1.3×10^{-208}
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
3	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
4	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
5	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 6: RPE procedure for the opisthokonta dataset, across all samples of gene trees. Candidate root partitions are sorted in descending order according to the significance values from the ROA-tests. The IDs of the sorted candidates correspond to the sequential ordering of the branches in Figure 5 (from leftmost to the rightmost branch). The RPE procedure was performed with and without pseudocounts (see text for details) and the inferred root partition(s) are shaded in green. The number of gene trees included in the RPE iterations is shown in the 'Sample size' column of the tables. N denotes the total number of gene families belonging to each of the gene family categories (CSC, CMC, PSC, PMC and combined). The 'Split frequency' column is the number of gene trees where the candidate root partition appears as a branch.

	With pseudocounts				Without pseudocounts			
	Partition ID	Split frequency	FDR adjusted p-value	Sample size	Partition ID	Split frequency	FDR adjusted p-value	Sample size
CSC N=117								
	1	116	-	-	1	116	-	-
	2	117	4.6×10^{-13}	117	2	117	7.0×10^{-14}	116
	3	117	3.9×10^{-21}	117	3	117	1.1×10^{-20}	117
	4	101	3.9×10^{-21}	117	4	101	1.6×10^{-18}	101
	5	115	3.9×10^{-21}	117	5	115	8.4×10^{-21}	115
CMC N=1328								
	1	957	-	-	1	957	-	-
	2	1040	4.1×10^{-45}	1119	2	1040	2.7×10^{-123}	878
	3	964	4.8×10^{-135}	1184	3	964	2.6×10^{-128}	899
	4	703	3.9×10^{-193}	1195	4	703	2.5×10^{-114}	692
	5	847	3.9×10^{-186}	1202	5	847	5.6×10^{-132}	840
PSC N=10085								
	1	418	-	-	1	418	-	-
	3	587	2.5×10^{-29}	794	2	650	3.7×10^{-26}	235
	2	650	3.3×10^{-24}	1206	3	587	1.8×10^{-33}	214
	4	1012	1.2×10^{-72}	1730	4	1012	1.7×10^{-45}	488
	5	780	2.4×10^{-37}	1730	5	780	8.7×10^{-33}	780
PMC N=6928								
	1	1295	-	-	1	1295	-	-
	3	2129	1.2×10^{-74}	2651	2	1087	1.3×10^{-84}	718
	2	1087	1.6×10^{-44}	2954	3	2129	5.9×10^{-119}	839
	4	2300	2.2×10^{-277}	3640	4	2300	3.9×10^{-153}	1614
	5	1356	4.2×10^{-150}	3640	5	1356	1.6×10^{-95}	1356
Combined N=18458								
	1	2786	-	-	1	2786	-	-
	2	2894	7.1×10^{-91}	3733	2	2894	6.0×10^{-242}	1947
	3	3797	1.8×10^{-299}	5461	3	3797	8.7×10^{-296}	2069
	4	4116	$<10^{-323}$	6682	4	4116	$<10^{-323}$	2895
	5	3098	$<10^{-323}$	6689	5	3098	1.1×10^{-274}	3091

Cyanobacteria

The cyanobacteria dataset supplies a different type of a positive control for our tests. In comparison to the opisthokonta dataset, the cyanobacteria dataset poses three types of difficulties for root inference: 1) deep phylogenetic relations among the species, 2) a large species set (130 species), and 3) inference of the root in the presence of LGT.

The majority-rule approach recovers the previously reported cyanobacteria root as the consensus root from the CSC gene trees. For this dataset, 70% of the CSC gene trees have the inferred MAD root on a branch separating 31 unicellular species (SynProCya clade) from the others, including multicellular and unicellular species. Notably, the known root partition was the only candidate with a significant result from the ROA-tests, across all but one sample of gene trees (CSC, CMC, PSC, and all gene trees combined) (Table 7, Figure 9).

The application of RPE procedure to the cyanobacteria dataset recovered the known cyanobacteria root as a single root inference in all but one sample of gene trees (i.e., CSC, PSC, PMC, and all trees combined). For the CMC gene trees the RPE procedure infers a root neighborhood that includes the known root partition and one additional candidate. However, the apparent root neighborhood obtained with this gene tree sample is due to the absence of CMC gene trees to perform the test in the last iteration of the RPE (Table 8). Consequently, the RPE procedure halts prematurely inferring one apparent root neighborhood of two root partitions.

Our root inference for the cyanobacteria dataset is in agreement with our previous finding from Section 4. The root partition implies that the cyanobacterial LCA was unicellular and that evolution of multicellularity in the phylum was a later innovation.

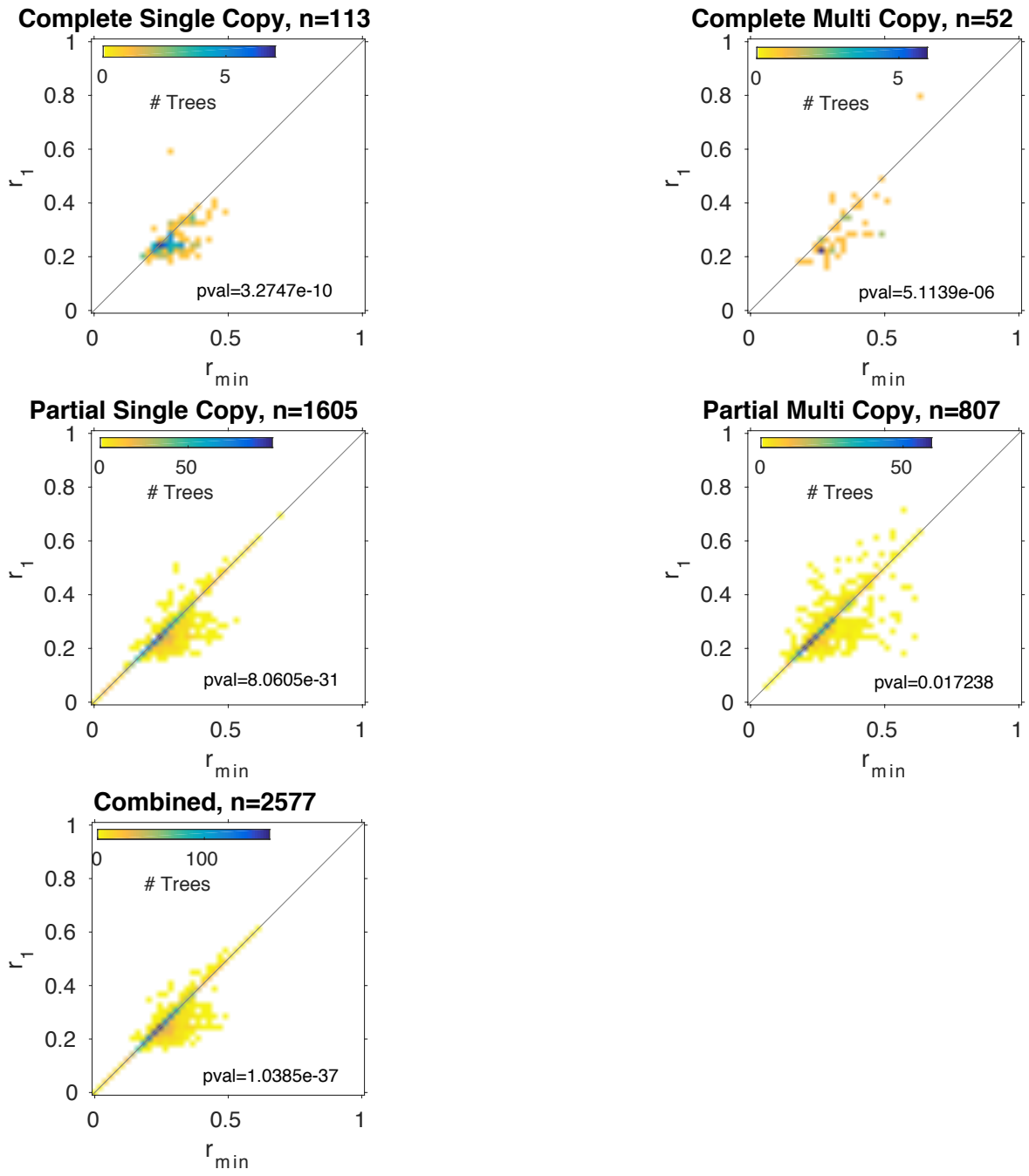


Figure 9: ROA-test for the consensus root branch in the cyanobacteria dataset, across all gene family categories.
See Figure 8 for layout details.

Table 7: ROA-test for all candidate root partitions in the cyanobacteria dataset. The tests were performed across all samples of gene trees. The nominal *p*-values of the tests are indicated and the cells shaded in green show significant tests (*p*-value<0.05).

	With pseudocounts					Without pseudocounts				
	CSC	CMC	PSC	PMC	Combined	CSC	CMC	PSC	PMC	Combined
1	6.1×10^{-9}	2.8×10^{-4}	2.6×10^{-10}	1.0	2.8×10^{-4}	3.3×10^{-10}	5.1×10^{-6}	8.1×10^{-31}	0.017	1.0×10^{-37}
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
3	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
4	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
5	1.0	1.0	1.0	1.0	1.0	0.25	0.50	1.0	1.0	1.0
6	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
7	1.0	-	1.0	1.0	1.0	0.50	-	1.0	1.0	1.0
8	1.0	-	1.0	1.0	1.0	0.50	-	1.0	1.0	1.0
9	1.0	-	1.0	1.0	1.0	0.50	-	1.0	1.0	1.0
10	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
11	1.0	-	1.0	1.0	1.0	0.75	-	1.0	1.0	1.0
12	1.0	-	1.0	1.0	1.0	0.50	-	1.0	1.0	1.0
13	1.0	1.0	1.0	1.0	1.0	0.88	0.50	1.0	1.0	1.0
14	1.0	-	1.0	1.0	1.0	0.75	-	1.0	1.0	1.0
15	1.0	1.0	1.0	1.0	1.0	0.88	1.0	1.0	1.0	1.0
16	1.0	-	1.0	1.0	1.0	0.75	-	1.0	1.0	1.0
17	1.0	1.0	1.0	1.0	1.0	1.0	0.97	1.0	1.0	1.0
18	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
19	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 8: RPE procedure for the cyanobacteria dataset, across all samples of gene trees.
See Table 6 for layout details.

	With pseudocounts				Without pseudocounts			
	Partition ID	Split frequency	FDR adjusted p-value	Sample size	Partition ID	Split frequency	FDR adjusted p-value	Sample size
CSC N=115								
	1	113	-	-	1	113	-	-
	2	58	1.3×10^{-16}	113	5	2	-	-
	3	101	8.3×10^{-19}	114	7	1	-	-
	4	16	1.5×10^{-20}	114	8	1	-	-
	5	2	4.5×10^{-19}	115	9	1	-	-
	6	113	1.0×10^{-20}	115	12	1	-	-
	7	1	7.6×10^{-21}	115	11	2	-	-
	8	1	7.6×10^{-21}	115	14	2	-	-
	9	1	7.6×10^{-21}	115	16	2	-	-
	10	11	7.6×10^{-21}	115	13	3	-	-
	11	2	7.6×10^{-21}	115	15	3	0.88	3
	12	1	7.6×10^{-21}	115	4	16	0.0028	16
	13	3	7.6×10^{-21}	115	10	11	9.8×10^{-4}	11
	14	2	7.6×10^{-21}	115	17	33	7.7×10^{-6}	33
	15	3	7.6×10^{-21}	115	2	58	5.8×10^{-7}	58
	16	2	7.6×10^{-21}	115	18	39	3.0×10^{-8}	39
	17	33	7.6×10^{-21}	115	19	54	1.3×10^{-10}	53
	18	39	7.0×10^{-21}	115	3	101	4.7×10^{-18}	101
	19	54	7.0×10^{-21}	115	6	113	1.5×10^{-20}	113
CMC N=57								
	1	53	-	-	1	53	-	-
	5	1	-	0	5	1	-	-
	2	25	1.0×10^{-8}	54	13	1	-	-
	3	40	2.3×10^{-9}	55	17	5	-	-
	6	49	9.3×10^{-11}	57	2	25	-	-
	13	1	2.8×10^{-11}	57	18	17	-	-
	17	5	2.8×10^{-11}	57	19	25	-	-
	4	5	2.6×10^{-11}	57	3	40	-	-
	10	2	2.6×10^{-11}	57	6	49	-	-
	15	1	2.6×10^{-11}	57	4	5	-	-
	18	17	2.6×10^{-11}	57	10	2	-	-
	19	25	2.6×10^{-11}	57	15	1	0.50	1
	7	0	-	0	7	0	-	0
	8	0	-	0	8	0	-	0
	9	0	-	0	9	0	-	0
	11	0	-	0	11	0	-	0
	12	0	-	0	12	0	-	0
	14	0	-	0	14	0	-	0
	16	0	-	0	16	0	-	0

	With pseudocounts				Without pseudocounts			
	Partition ID	Split frequency	FDR adjusted p-value	Sample size	Partition ID	Split frequency	FDR adjusted p-value	Sample size
PSC N=16285								
	1	1607	-	-	1	1607	-	-
	2	1613	3.6×10^{-103}	2065	2	1613	1.4×10^{-26}	1155
	3	1885	6.1×10^{-133}	3039	3	1885	7.3×10^{-104}	911
	4	1985	1.6×10^{-172}	3039	4	1985	5.2×10^{-30}	1985
	5	2555	2.0×10^{-251}	4510	5	2555	3.2×10^{-37}	1084
	6	2122	4.6×10^{-147}	5102	6	2122	1.1×10^{-164}	1530
	7	4362	$<10^{-323}$	7900	7	4362	5.5×10^{-9}	1564
	8	4317	$<10^{-323}$	9622	8	4317	4.1×10^{-19}	2595
	9	2636	2.2×10^{-284}	9954	9	2636	1.6×10^{-34}	2304
	10	1939	7.1×10^{-260}	9954	10	1939	2.1×10^{-75}	1939
	11	2136	1.0×10^{-216}	10127	11	2136	2.7×10^{-68}	1963
	12	3064	$<10^{-323}$	10440	12	3064	1.5×10^{-41}	2751
	13	2023	4.9×10^{-277}	10440	13	2023	7.1×10^{-126}	2023
	14	2527	$<10^{-323}$	10483	14	2527	3.1×10^{-99}	2484
	15	3180	$<10^{-323}$	10486	15	3180	7.0×10^{-79}	3177
	16	1799	2.4×10^{-286}	10606	16	1799	7.0×10^{-69}	1679
	17	1962	1.9×10^{-256}	10624	17	1962	1.5×10^{-103}	1944
	18	1447	9.7×10^{-178}	11067	18	1447	8.8×10^{-69}	1004
	19	842	4.1×10^{-117}	11313	19	842	4.4×10^{-43}	596
PMC N=4518								
	1	811	-	-	1	811	-	-
	2	754	1.3×10^{-48}	976	7	766	0.014	31
	3	972	7.3×10^{-52}	1541	2	754	3.3×10^{-8}	596
	4	956	6.7×10^{-101}	1541	3	972	2.8×10^{-33}	469
	5	1087	7.6×10^{-151}	2146	4	956	8.7×10^{-17}	956
	6	1088	1.5×10^{-45}	2438	5	1087	2.3×10^{-12}	605
	7	766	$<10^{-323}$	2937	6	1088	5.6×10^{-67}	840
	8	788	$<10^{-323}$	3271	8	788	0.0075	454
	9	1053	2.0×10^{-181}	3422	9	1053	4.1×10^{-12}	902
	10	905	3.6×10^{-154}	3422	10	905	5.0×10^{-35}	905
	11	980	6.5×10^{-152}	3540	11	980	2.1×10^{-44}	862
	12	1019	3.6×10^{-290}	3675	12	1019	7.0×10^{-18}	884
	13	940	1.0×10^{-165}	3675	13	940	1.6×10^{-60}	940
	14	1012	2.1×10^{-218}	3737	14	1012	5.0×10^{-27}	950
	15	1170	$<10^{-323}$	3746	15	1170	1.4×10^{-37}	1161
	16	837	2.6×10^{-156}	3775	16	837	1.5×10^{-32}	808
	17	961	7.1×10^{-140}	3788	17	961	9.1×10^{-36}	948
	18	567	8.3×10^{-100}	3824	18	567	1.1×10^{-37}	531
	19	216	3.2×10^{-58}	3847	19	216	9.0×10^{-27}	193

	With pseudocounts				Without pseudocounts			
	Partition ID	Split frequency	FDR adjusted p-value	Sample size	Partition ID	Split frequency	FDR adjusted p-value	Sample size
Combined N=20975								
	1	2584	-	-	1	2584	-	-
	2	2450	6.9×10^{-171}	3207	2	2450	1.0×10^{-36}	1827
	3	2998	3.2×10^{-202}	4749	3	2998	6.5×10^{-155}	1456
	4	2962	9.1×10^{-299}	4749	4	2962	5.9×10^{-46}	2962
	5	3645	$<10^{-323}$	6826	5	3645	2.4×10^{-49}	1568
	6	3372	3.0×10^{-212}	7712	6	3372	3.4×10^{-263}	2486
	7	5129	$<10^{-323}$	11009	7	5129	2.3×10^{-9}	1832
	8	5106	$<10^{-323}$	13065	8	5106	6.1×10^{-20}	3050
	9	3690	$<10^{-323}$	13548	9	3690	1.4×10^{-44}	3207
	10	2857	$<10^{-323}$	13548	10	2857	1.0×10^{-110}	2857
	11	3118	$<10^{-323}$	13839	11	3118	8.7×10^{-111}	2827
	12	4084	$<10^{-323}$	14287	12	4084	3.2×10^{-57}	3636
	13	2967	$<10^{-323}$	14287	13	2967	2.1×10^{-185}	2967
	14	3541	$<10^{-323}$	14392	14	3541	1.2×10^{-122}	3436
	15	4354	$<10^{-323}$	14404	15	4354	5.2×10^{-115}	4342
	16	2638	$<10^{-323}$	14553	16	2638	4.8×10^{-99}	2489
	17	2961	$<10^{-323}$	14584	17	2961	1.4×10^{-140}	2930
	18	2070	1.6×10^{-303}	15063	18	2070	7.1×10^{-114}	1591
	19	1137	1.7×10^{-202}	15332	19	1137	5.7×10^{-80}	868

Proteobacteria

In the proteobacteria dataset we encountered an uncertain consensus root with the majority-rule approach, contrasting the results obtained with the eukaryotic and cyanobacteria datasets. The consensus root branch separates epsilon-proteobacteria from the other classes (Figure 5). However, this branch accounts for only 25% of the root inferences from CSC gene trees. The second most frequent root is a branch joining epsilonproteobacteria to deltaproteobacteria, accounting for 18% of the root inferences. Notably, the uncertainty in this datasets accompanies smaller sample size in comparison to the eukaryotic and cyanobacteria datasets (Figure 5 and Table 4). The same root uncertainty is observed with the ROA-tests (Figure 10). None of the root candidates yield a significant ROA-test. In fact, most of *p*-values are close to one, indicating a tight competition among the candidates (Table 9). In such situation, the RPE procedure is required for either disambiguation among competing candidates or inference of a root neighborhood. The RPE procedure disambiguated the proteobacteria root, with the partition that separates epsilonproteobacteria from the other classes as the best candidate. The RPE consistently recovered the epsilon root partition across all the samples of gene trees, and achieved maximal significance when information from all the gene trees was combined. Also in this dataset, the use of

pseudocounts assisted to counteract biases stemming from small sample sizes during the RPE procedure (Table 10).

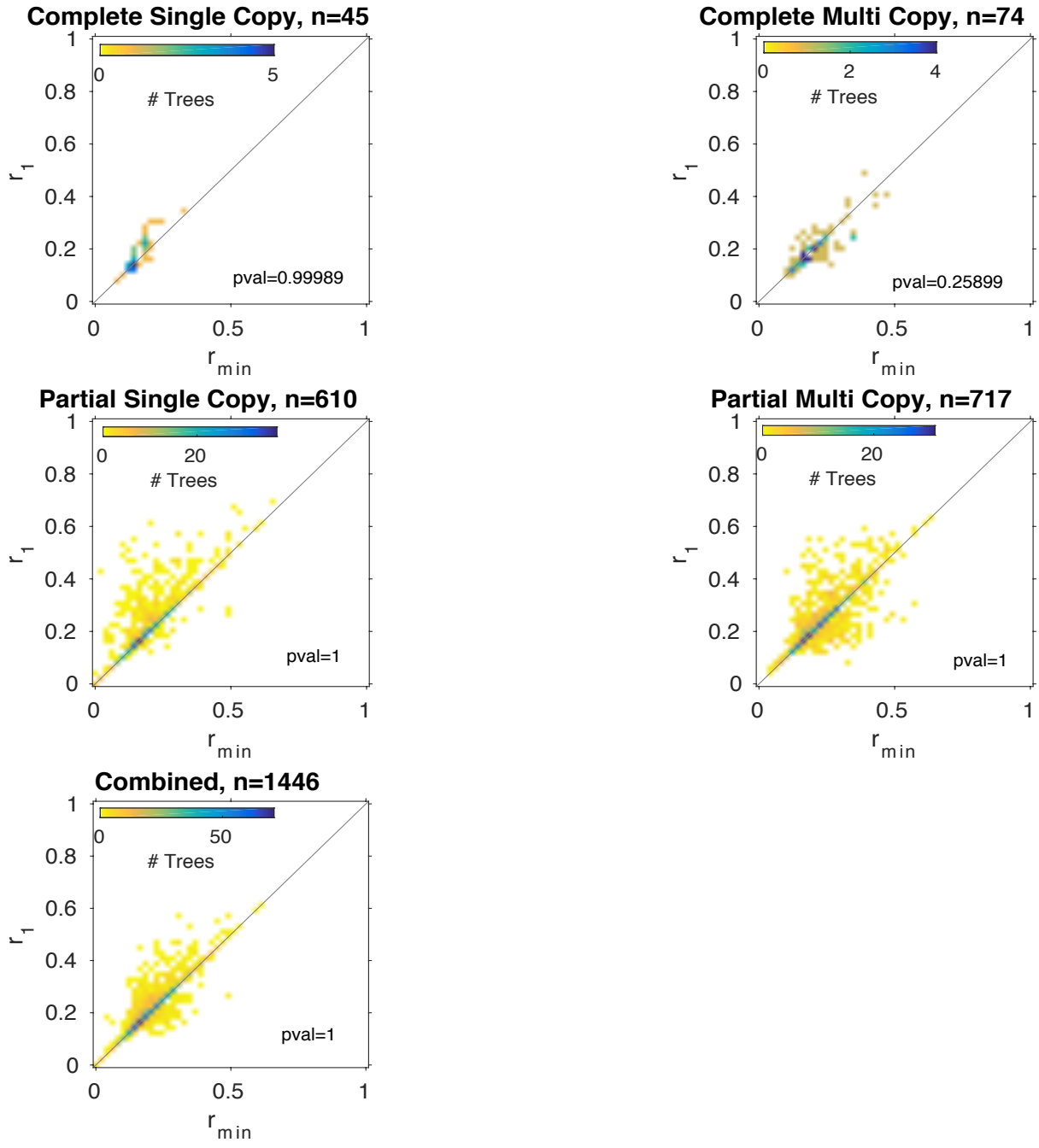


Figure 10: ROA-test for the consensus root branch in the proteobacteria dataset, across all gene family categories.
See Figure 8 for layout details.

Table 9: ROA-test for all candidate root partitions in the proteobacteria dataset.
The tests were performed across all samples of gene trees. The nominal *p*-values of the tests are indicated.

	With pseudocounts					Without pseudocounts				
	CSC	CMC	PSC	PMC	Combined	CSC	CMC	PSC	PMC	Combined
1	1.0	0.21	1.0	1.0	1.0	1.0	0.26	1.0	1.0	1.0
2	1.0	1.0	1.0	1.0	1.0	1.0	0.98	1.0	1.0	1.0
3	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
4	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
5	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
6	1.0	-	1.0	1.0	1.0	0.63	-	1.0	1.0	1.0
7	1.0	1.0	1.0	1.0	1.0	0.25	1.0	1.0	1.0	1.0
8	1.0	1.0	1.0	1.0	1.0	0.81	1.0	1.0	1.0	1.0
9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
10	1.0	-	1.0	1.0	1.0	0.50	-	1.0	1.0	1.0
11	1.0	1.0	1.0	1.0	1.0	0.81	0.50	1.0	1.0	1.0
12	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
13	1.0	-	1.0	1.0	1.0	0.50	-	1.0	1.0	1.0
14	1.0	-	1.0	1.0	1.0	0.50	-	1.0	1.0	1.0
15	1.0	-	1.0	1.0	1.0	0.50	-	1.0	1.0	1.0
16	1.0	-	1.0	1.0	1.0	0.50	-	1.0	1.0	1.0
17	1.0	1.0	1.0	1.0	1.0	0.50	1.0	1.0	1.0	1.0
18	1.0	-	1.0	1.0	1.0	0.50	-	1.0	1.0	1.0
19	1.0	-	1.0	1.0	1.0	0.88	-	1.0	1.0	1.0
20	1.0	-	1.0	1.0	1.0	0.50	-	1.0	1.0	1.0

Table 10: RPE procedure for the proteobacteria dataset, across all samples of gene trees.
See Table 6 for layout details.

	With pseudocounts				Without pseudocounts			
	Partition ID	Split frequency	FDR adjusted p-value	Sample size	Partition ID	Split frequency	FDR adjusted p-value	Sample size
CSC N=45								
	1	45	-	-	7	2	-	-
	5	34	0.0074	45	10	1	-	-
	2	23	1.1×10^{-6}	45	13	2	-	-
	3	44	0.0014	45	14	1	-	-
	4	37	7.3×10^{-9}	45	15	1	-	-
	9	14	4.3×10^{-9}	45	16	1	-	-
	6	3	3.3×10^{-9}	45	17	1	-	-
	7	2	3.3×10^{-9}	45	18	1	-	-
	8	4	3.3×10^{-9}	45	20	1	-	-
	11	4	3.3×10^{-9}	45	6	3	-	-
	13	2	3.3×10^{-9}	45	8	4	-	-
	10	1	2.9×10^{-9}	45	11	4	-	-
	12	45	2.9×10^{-9}	45	19	3	-	-
	14	1	2.9×10^{-9}	45	2	23	-	-
	15	1	2.9×10^{-9}	45	9	14	0.25	12
	16	1	2.9×10^{-9}	45	1	45	4.6×10^{-4}	42
	17	1	2.9×10^{-9}	45	5	34	6.4×10^{-5}	34
	18	1	2.9×10^{-9}	45	4	37	1.1×10^{-6}	37
	19	3	2.9×10^{-9}	45	3	44	4.4×10^{-8}	44
	20	1	2.9×10^{-9}	45	12	45	2.9×10^{-9}	45
CMC N=85								
	1	76	-	-	1	76	-	-
	3	53	5.7×10^{-7}	76	11	1	-	-
	12	66	1.2×10^{-12}	77	2	16	-	-
	4	49	5.1×10^{-14}	77	9	16	-	-
	5	31	5.1×10^{-14}	77	3	53	-	-
	2	16	4.7×10^{-14}	77	5	31	-	-
	9	16	1.8×10^{-14}	77	4	49	-	-
	11	1	1.3×10^{-14}	77	12	66	-	-
	7	2	1.3×10^{-14}	77	7	2	-	-
	8	1	1.3×10^{-14}	77	8	1	-	-
	17	2	1.3×10^{-14}	77	17	2	0.25	2
	6	0	-	0	6	0	-	0
	10	0	-	0	10	0	-	0
	13	0	-	0	13	0	-	0
	14	0	-	0	14	0	-	0
	15	0	-	0	15	0	-	0
	16	0	-	0	16	0	-	0
	18	0	-	0	18	0	-	0
	19	0	-	0	19	0	-	0
	20	0	-	0	20	0	-	0

	With pseudocounts				Without pseudocounts			
	Partition ID	Split frequency	FDR adjusted p-value	Sample size	Partition ID	Split frequency	FDR adjusted p-value	Sample size
PSC N=8496								
	1	610	-	-	1	610	-	-
	2	2340	6.1×10^{-34}	2680	2	2340	0.91	270
	3	2481	8.8×10^{-8}	4048	3	2481	0.0040	1113
	4	1475	1.6×10^{-19}	4757	4	1475	1.3×10^{-30}	766
	5	2910	2.4×10^{-171}	4802	5	2910	5.6×10^{-47}	2865
	6	2457	1.5×10^{-179}	5140	6	2457	7.1×10^{-56}	2119
	7	2937	1.5×10^{-219}	5199	7	2937	6.5×10^{-68}	2878
	8	2954	3.4×10^{-188}	5447	8	2954	7.3×10^{-44}	2706
	9	2324	1.5×10^{-208}	5449	9	2324	7.0×10^{-83}	2322
	10	2931	1.4×10^{-271}	5551	10	2931	6.9×10^{-102}	2829
	11	2361	2.9×10^{-226}	5560	11	2361	4.5×10^{-99}	2352
	12	955	7.9×10^{-37}	5686	12	955	9.8×10^{-59}	829
	13	2856	8.1×10^{-254}	5686	13	2856	2.8×10^{-131}	2856
	14	2573	2.8×10^{-259}	5721	14	2573	8.1×10^{-104}	2538
	15	3133	1.7×10^{-302}	5856	15	3133	2.4×10^{-78}	2998
	16	2023	3.6×10^{-205}	6288	16	2023	2.8×10^{-137}	1591
	17	1999	9.4×10^{-186}	6307	17	1999	2.5×10^{-48}	1980
	18	3103	$<10^{-323}$	6316	18	3103	4.3×10^{-109}	3094
	19	956	2.6×10^{-108}	6321	19	956	1.1×10^{-74}	951
	20	2702	1.4×10^{-265}	6324	20	2702	4.0×10^{-116}	2699
PMC N=4835								
	1	778	-	-	1	778	-	-
	2	1045	5.1×10^{-92}	1680	2	1045	0.86	143
	3	1224	6.0×10^{-43}	2410	3	1224	7.2×10^{-10}	494
	4	962	8.8×10^{-24}	2847	4	962	4.0×10^{-23}	525
	5	1130	1.4×10^{-234}	2898	5	1130	1.4×10^{-14}	1079
	6	1023	1.1×10^{-207}	3042	6	1023	7.1×10^{-19}	879
	7	1105	7.3×10^{-275}	3060	7	1105	2.7×10^{-27}	1087
	8	1194	4.8×10^{-261}	3146	8	1194	6.8×10^{-30}	1108
	9	1050	3.8×10^{-249}	3153	9	1050	2.1×10^{-25}	1043
	10	1050	7.0×10^{-306}	3169	10	1050	1.0×10^{-40}	1034
	11	1019	1.3×10^{-257}	3181	11	1019	5.4×10^{-39}	1007
	12	915	1.5×10^{-22}	3413	12	915	8.1×10^{-48}	683
	13	1225	$<10^{-323}$	3413	13	1225	5.1×10^{-60}	1225
	14	1033	1.6×10^{-302}	3440	14	1033	1.4×10^{-30}	1006
	15	1059	$<10^{-323}$	3485	15	1059	2.4×10^{-24}	1014
	16	1091	7.7×10^{-193}	3697	16	1091	1.5×10^{-64}	879
	17	1006	1.2×10^{-237}	3770	17	1006	3.0×10^{-11}	933
	18	1124	$<10^{-323}$	3775	18	1124	3.3×10^{-31}	1119
	19	748	1.7×10^{-125}	3810	19	748	2.4×10^{-33}	713
	20	1146	$<10^{-323}$	3810	20	1146	1.3×10^{-55}	1146

	With pseudocounts				Without pseudocounts			
	Partition ID	Split frequency	FDR adjusted p-value	Sample size	Partition ID	Split frequency	FDR adjusted p-value	Sample size
Combined N=13461								
	1	1509	-	-	1	1509	-	-
	2	3424	3.5×10^{-141}	4481	2	3424	0.97	452
	3	3802	1.5×10^{-48}	6579	3	3802	1.1×10^{-9}	1704
	4	2523	6.8×10^{-52}	7725	4	2523	6.2×10^{-62}	1377
	5	4105	$<10^{-323}$	7821	5	4105	3.2×10^{-64}	4009
	6	3483	$<10^{-323}$	8303	6	3483	5.1×10^{-73}	3001
	7	4046	$<10^{-323}$	8380	7	4046	2.0×10^{-93}	3969
	8	4153	$<10^{-323}$	8714	8	4153	2.4×10^{-70}	3819
	9	3404	$<10^{-323}$	8723	9	3404	2.1×10^{-110}	3395
	10	3982	$<10^{-323}$	8841	10	3982	1.8×10^{-140}	3864
	11	3385	$<10^{-323}$	8862	11	3385	2.8×10^{-136}	3364
	12	1981	1.3×10^{-65}	9221	12	1981	2.8×10^{-119}	1622
	13	4083	$<10^{-323}$	9221	13	4083	7.9×10^{-190}	4083
	14	3607	$<10^{-323}$	9283	14	3607	1.2×10^{-132}	3545
	15	4193	$<10^{-323}$	9463	15	4193	1.6×10^{-100}	4013
	16	3115	$<10^{-323}$	10107	16	3115	6.3×10^{-200}	2471
	17	3008	$<10^{-323}$	10199	17	3008	7.2×10^{-57}	2916
	18	4228	$<10^{-323}$	10213	18	4228	2.2×10^{-138}	4214
	19	1707	4.2×10^{-253}	10253	19	1707	3.4×10^{-106}	1667
	20	3849	$<10^{-323}$	10256	20	3849	1.3×10^{-169}	3846

Archaea

In the archaea dataset the best root is even more uncertain than observed for the proteobacteria dataset (Figures 5). The consensus root branch from CSC trees separates halobacteria and methanomicrobia from other archaea species, but accounting for only 12% of the root inferences. For 9% of the trees, MAD infers the root on a branch that separates the single nanoarchaeum species from the other species. We note that sample size for the archaea dataset was the smallest among the demonstrative datasets (Figure 5). The root uncertainty of this dataset is also noticeable in the ROA-tests, for which no candidate attained a significant result (Table 11, Figure 11).

When we applied the RPE procedure to this dataset, the inference was not of a single root. Instead, the inference is of a root neighborhood comprising 5 root partitions: 3 partitions within the euryarchaeota phylum, one partition separating the nanoarchaeota species from the others and one partition joining nanoarchaeota with thermoprotei (crenarchaeota phylum) (Table 12). We observed the same neighborhood inference for all but one sample of gene trees. The inference with CSC gene trees alone renders a neighborhood of two partitions: the partition that joins halobacteria with methanomicrobia (both euryarchaeota) and the partition leading to nanoarchaeota. Note that the

neighborhood inferred with CSC trees alone is contained in the neighborhood obtained with the other tree samples. The difference in terms of number of partitions for the root neighborhood may stem from the extremely small sample of CSC trees for this dataset (see split frequency for CSC gene trees in Table 12).

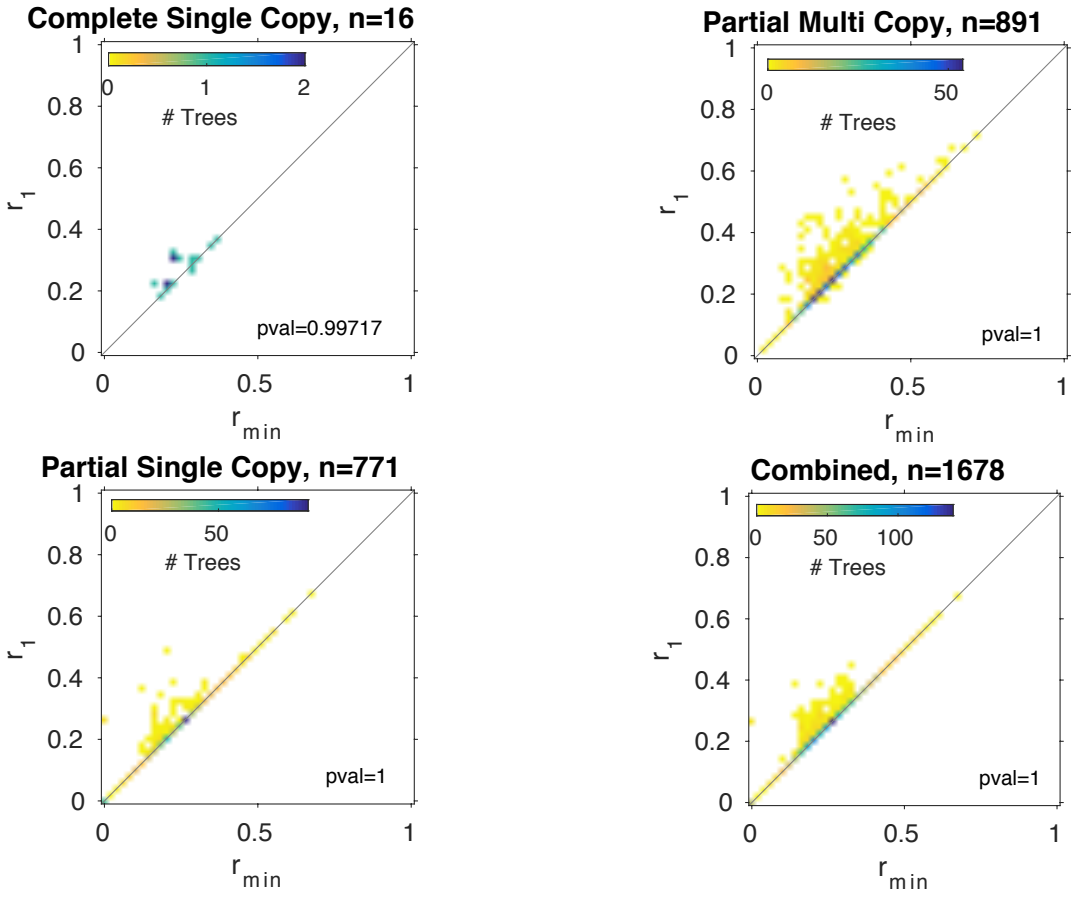


Figure 11: ROA-test for the consensus root branch in the archaea dataset, across all gene family categories.
See Figure 8 for layout details. Note that CMC gene trees are absent in this dataset.

Table 11: ROA-test for all candidate root partitions in the archaea dataset. The tests were performed across all samples of gene trees. The nominal *p*-values of the tests are indicated.

	With pseudocounts					Without pseudocounts				
	CSC	CMC	PSC	PMC	Combined	CSC	CMC	PSC	PMC	Combined
1	1.0	-	1.0	1.0	1.0	1.0	-	1.0	1.0	1.0
2	1.0	-	1.0	0.91	1.0	1.0	-	1.0	1.0	1.0
3	1.0	-	1.0	1.0	1.0	0.59	-	1.0	1.0	1.0
4	1.0	-	1.0	1.0	1.0	0.50	-	1.0	1.0	1.0
5	1.0	-	1.0	1.0	1.0	1.0	-	1.0	1.0	1.0
6	1.0	-	1.0	1.0	1.0	0.75	-	1.0	1.0	1.0
7	1.0	-	1.0	1.0	1.0	0.50	-	1.0	1.0	1.0
8	1.0	-	1.0	1.0	1.0	0.50	-	1.0	1.0	1.0
9	1.0	-	1.0	1.0	1.0	0.99	-	1.0	1.0	1.0
10	1.0	-	1.0	1.0	1.0	0.50	-	1.0	1.0	1.0
11	1.0	-	1.0	1.0	1.0	0.88	-	1.0	1.0	1.0
12	1.0	-	1.0	1.0	1.0	0.75	-	1.0	1.0	1.0
13	1.0	-	1.0	1.0	1.0	0.88	-	1.0	1.0	1.0
14	1.0	-	1.0	1.0	1.0	0.92	-	1.0	1.0	1.0
15	1.0	-	1.0	1.0	1.0	0.75	-	1.0	1.0	1.0
16	1.0	-	1.0	1.0	1.0	0.75	-	1.0	1.0	1.0
17	1.0	-	1.0	1.0	1.0	0.89	-	1.0	1.0	1.0
18	1.0	-	1.0	1.0	1.0	0.75	-	1.0	1.0	1.0
19	1.0	-	1.0	1.0	1.0	0.50	-	1.0	1.0	1.0
20	1.0	-	1.0	1.0	1.0	0.50	-	1.0	1.0	1.0
21	1.0	-	1.0	1.0	1.0	0.50	-	1.0	1.0	1.0
22	1.0	-	1.0	1.0	1.0	0.97	-	1.0	1.0	1.0

Table 12: RPE procedure for the archaea dataset, across all samples of gene trees.
See Table 6 for layout details.

	With pseudocounts				Without pseudocounts			
	Partition ID	Split frequency	FDR adjusted p-value	Sample size	Partition ID	Split frequency	FDR adjusted p-value	Sample size
CSC N=26								
	2	26	-	-	4	1	-	-
	1	16	0.010	26	7	1	-	-
	3	5	2.2×10^{-5}	26	8	1	-	-
	17	6	3.3×10^{-5}	26	10	2	-	-
	14	6	2.4×10^{-5}	26	19	1	-	-
	5	26	1.9×10^{-5}	26	20	1	-	-
	6	3	7.0×10^{-6}	26	21	1	-	-
	10	2	6.3×10^{-6}	26	3	5	-	-
	11	4	5.6×10^{-6}	26	6	3	-	-
	15	3	5.6×10^{-6}	26	12	2	-	-
	4	1	5.0×10^{-6}	26	15	3	-	-
	7	1	5.0×10^{-6}	26	16	2	-	-
	8	1	5.0×10^{-6}	26	18	2	-	-
	9	7	5.0×10^{-6}	26	11	4	-	-
	12	2	5.0×10^{-6}	26	13	3	-	-
	13	3	5.0×10^{-6}	26	17	6	-	-
	16	2	5.0×10^{-6}	26	14	6	-	-
	18	2	5.0×10^{-6}	26	22	5	-	-
	19	1	5.0×10^{-6}	26	9	7	0.16	6
	20	1	5.0×10^{-6}	26	1	16	0.030	16
	21	1	5.0×10^{-6}	26	2	26	4.6×10^{-5}	26
	22	5	5.0×10^{-6}	26	5	26	5.6×10^{-6}	26
PSC N=4860								
	2	64	-	-	2	64	-	-
	1	771	-	-	1	771	-	-
	3	490	-	-	3	490	0.45	183
	4	716	-	-	4	716	8.8×10^{-12}	510
	5	559	0.83	1507	5	559	2.6×10^{-5}	379
	6	753	6.2×10^{-56}	1517	6	753	1.4×10^{-20}	743
	7	805	3.2×10^{-64}	1526	7	805	2.9×10^{-20}	796
	8	830	2.1×10^{-61}	1574	8	830	6.7×10^{-15}	782
	9	798	7.4×10^{-72}	1575	9	798	2.2×10^{-24}	797
	10	738	1.0×10^{-70}	1609	10	738	2.0×10^{-19}	704
	11	319	3.9×10^{-30}	1731	11	319	1.2×10^{-11}	197
	12	476	1.7×10^{-48}	1852	12	476	4.0×10^{-10}	355
	13	505	9.8×10^{-47}	1887	13	505	2.0×10^{-12}	470
	14	490	3.0×10^{-55}	1887	14	490	1.8×10^{-23}	490
	15	294	8.5×10^{-28}	1897	15	294	6.4×10^{-16}	284
	16	546	6.2×10^{-62}	1898	16	546	2.2×10^{-19}	545
	17	526	2.6×10^{-61}	1898	17	526	9.4×10^{-26}	526
	18	292	1.7×10^{-38}	1898	18	292	2.3×10^{-28}	292
	19	340	8.0×10^{-36}	1901	19	340	2.3×10^{-17}	337
	20	546	1.7×10^{-47}	1901	20	546	9.4×10^{-26}	546
	21	434	2.4×10^{-57}	1912	21	434	1.0×10^{-19}	423

	22	527	3.9×10^{-66}	1912	22	527	9.9×10^{-29}	527
	With pseudocounts				Without pseudocounts			
	Partition ID	Split frequency	FDR adjusted p-value	Sample size	Partition ID	Split frequency	FDR adjusted p-value	Sample size
PMC N=4826								
	2	266	-	-	2	266	-	-
	1	891	-	-	1	891	-	-
	3	649	-	-	3	649	0.41	220
	4	829	-	-	4	829	4.7×10^{-20}	504
	5	1165	0.83	2291	5	1165	3.9×10^{-4}	727
	6	838	1.0×10^{-196}	2302	6	838	3.5×10^{-22}	827
	7	812	6.5×10^{-207}	2337	7	812	1.6×10^{-18}	777
	8	801	1.6×10^{-209}	2370	8	801	1.1×10^{-10}	768
	9	817	3.5×10^{-221}	2373	9	817	9.9×10^{-25}	814
	10	754	3.1×10^{-202}	2421	10	754	3.5×10^{-19}	706
	11	531	1.2×10^{-58}	2588	11	531	5.5×10^{-19}	364
	12	523	1.3×10^{-165}	2685	12	523	5.5×10^{-22}	426
	13	637	5.1×10^{-185}	2713	13	637	5.7×10^{-22}	609
	14	683	6.0×10^{-187}	2714	14	683	3.4×10^{-30}	682
	15	662	2.1×10^{-69}	2825	15	662	4.4×10^{-11}	551
	16	695	4.7×10^{-233}	2831	16	695	3.2×10^{-23}	689
	17	711	1.4×10^{-223}	2831	17	711	7.9×10^{-29}	711
	18	685	2.6×10^{-128}	2832	18	685	2.9×10^{-46}	684
	19	681	4.6×10^{-123}	2882	19	681	1.8×10^{-23}	631
	20	904	1.8×10^{-204}	2887	20	904	1.7×10^{-43}	899
	21	570	1.9×10^{-202}	2903	21	570	3.4×10^{-33}	554
	22	689	5.1×10^{-238}	2904	22	689	2.9×10^{-39}	688
Combined N=9712								
	2	356	-	-	2	356	-	-
	1	1678	-	-	1	1678	-	-
	3	1144	-	-	3	1144	0.45	408
	4	1546	-	-	4	1546	1.8×10^{-29}	1015
	5	1750	0.88	3824	5	1750	1.4×10^{-8}	1132
	6	1594	7.3×10^{-255}	3845	6	1594	7.0×10^{-41}	1573
	7	1618	5.1×10^{-274}	3889	7	1618	3.3×10^{-37}	1574
	8	1632	2.4×10^{-274}	3970	8	1632	1.0×10^{-23}	1551
	9	1622	1.3×10^{-295}	3974	9	1622	4.5×10^{-47}	1618
	10	1494	5.2×10^{-277}	4056	10	1494	4.8×10^{-37}	1412
	11	854	2.9×10^{-90}	4345	11	854	6.9×10^{-29}	565
	12	1001	4.1×10^{-215}	4563	12	1001	1.3×10^{-29}	783
	13	1145	1.3×10^{-233}	4626	13	1145	1.2×10^{-31}	1082
	14	1179	5.5×10^{-244}	4627	14	1179	4.0×10^{-51}	1178
	15	959	8.5×10^{-98}	4748	15	959	1.0×10^{-23}	838
	16	1243	5.5×10^{-298}	4755	16	1243	1.1×10^{-40}	1236
	17	1243	6.4×10^{-287}	4755	17	1243	2.9×10^{-53}	1243
	18	979	2.2×10^{-168}	4756	18	979	1.4×10^{-73}	978
	19	1022	4.3×10^{-161}	4809	19	1022	2.7×10^{-38}	969
	20	1451	5.2×10^{-254}	4814	20	1451	1.8×10^{-67}	1446
	21	1005	7.3×10^{-262}	4841	21	1005	1.1×10^{-50}	978
	22	1221	4.9×10^{-306}	4842	22	1221	1.1×10^{-66}	1220

5.3 Conclusions

The MAD rooting method evaluates all the branches in a gene tree as a possible root position, generating an *r* value for each branch in the tree. The application of MAD to a sample of gene trees allows testing for differences in the magnitude of *r* values for alternative root partitions of a species set. Testing the distribution of *r* values does not require the inference of a strict root position for the individual gene trees. Here we built upon this property a series of statistical tests to infer the best root partition of a species set. The tests take into consideration a sample of unrooted gene trees and evaluates all candidate root partitions as a possible root.

We present three tests for the selection of the best root partition; each test has application in different analytical contexts. For species sets with no prior information regarding the root partition, the definition of a pool of root candidates from the CSC gene trees is followed by the use of the ROA-tests. The RPE procedure should be subsequently used in case of no clear best candidate, as judged by the significance values from ROA-tests. When prior information regarding the root partition is available (e.g., conflicting root partitions reported in the literature), the RPW-test is sufficient for testing two hypothesized root partitions. When the goal is to test simultaneously three or more hypothesized root partitions, the ROA-test is preferred. The ROA-test requires fewer tests for the evaluation of all the root partitions and it provides a more conservative statistical setting in comparison to the RPW-test. The RPE procedure may be subsequently necessary if no clear best root partition can be selected from the ROA-tests.

Our approach allows for the consideration of non-CSC gene trees – in addition to CSC gene trees – for root partition inferences. Our results demonstrate that in datasets with a strong root signal, the root inference with non-CSC gene trees reproduces the result obtained with CSC gene trees. This is exemplified in the opisthokonta and cyanobacteria datasets. The combination of all gene trees for root inferences increases the power of the test due to a considerable increase in sample size.

The consideration of non-CSC gene trees has a general implication in root partition inferences, since the prior definition of an ideal sample size is elusive. In the most extreme cases, the consideration of non-CSC gene trees paves the way for root inferences in datasets with no complete gene families. In the literature the total number of complete gene families is commonly termed as the ‘core genome’ of the ‘pan genome’, where the pan genome comprises all gene families in the species set. Notably, the size of the core genome decays with the increase of the number genomes in the species set (e.g., (Medini et al. 2005)). Previous studies suggested that nearly complete gene families should be included in the core genome (Puigbò et al. 2009). Nonetheless, the definition of nearly complete gene families may be debated. In this context, our approach permits the use of pan-genomes for root inferences, without committing to an arbitrary

definition of nearly complete gene families. Our approach utilized 10-17% of the total number of gene families for root partition inferences. This is in stark contrast to the 0.1-0.6% of the gene families utilized by CSC-based approaches, like the majority-rule. The number of genes families considered in our tests corresponds to the number of genes encoded in modern genomes, supplying 'total evidence' for root partition inferences.

Our analyses of the demonstrative datasets show that different species sets present varying levels of root signal: the opitsthokonta and cyanobacteria datasets show a strong root signal, the proteobacteria dataset has a moderate root signal and the root signal within the archaea is weak. Datasets with weak root signal are better described in terms of root neighborhoods, when two or more candidate root partitions attain similar *r* values in the sample of gene trees. This case is exemplified with the archaea dataset. Root neighborhoods may originate from methodological artifacts, that is, uncertainty in the root partition inference, alignments and tree reconstruction artifacts. Alternatively, root neighborhoods may stem from evolutionary scenarios that deviate from the bifurcating species tree framework. One such evolutionary scenario, for instance, is when a large number of genes trace back to different LCAs, i.e., reticulated evolution. Reticulated evolution may lead to the real existence of multiple root partitions for the species set. One root partition represents the organismal (vertical) LCA while the others correspond to the different donor lineages that contributed substantial genetic content via LGT. Another possible scenario is a rapid diversification of the organismal LCA into multiple lineages, resulting in a polytomous root branch in the underlying species tree. Accounting for root neighborhoods is possible because our approach neither requires nor assumes the existence of a bifurcating species tree. This is in contrast to the existing phylogenomic approaches that are confined within the framework of bifurcation of ancestral lineages.

LGT plays a major role in prokaryotic evolution, generating gene trees with contradicting evolutionary histories (Dagan et al. 2008). Notably, LGT among species within the same side of the true root partition will not bias the root inferences with our approach. Nonetheless, LGT events between species located in different sides of the root partition generate gene trees with signals contradicting the true root partition. Those LGTs are a source of noise for root inferences. LGT from unsampled species to recipients in the species set has the potential to trace back to a LCA that differs from the LCA of vertically inherited genes, becoming a potential source of noise for the root inferences. Yet more noise may come from ancient gene duplications followed by differential loss. Duplications followed by differential loss may give rise to gene trees that appear to trace back to different LCAs. In our approach we accommodate contradicting signals introduced by LGT, duplications and differential loss by considering the information from all informative gene families simultaneously. The consideration of multiple gene families enables the detection of the true root

partition even in face of sporadic non-vertical events (e.g., LGT). When non-vertical events are pervasive, however, the root inference appears as a root neighborhood.

A fundamental element in our approach is the prior definition of a pool of candidate root partitions. Spurious candidates, appearing in low frequency as a branch in the gene trees sample may drastically decrease the number of considered gene trees in the tests, resulting in apparent large root neighborhoods. Such spurious root candidates may originate from sporadic non-vertical events underlying the evolution of the gene family or, alternatively, tree reconstruction artifacts. We found out that the use of pseudocounts for root candidates missing in gene trees is useful in order to alleviate the drastic reduction of gene trees considered in the tests, as exemplified in the RPE procedure for the cyanobacteria dataset. Another possibility is the exclusion of spurious root candidates from the pool. One criterion for the exclusion, for example, could be quality of the CSC gene trees where the candidates appear as a root branch. Alternatively, one could exclude root candidates having low branch frequency in the total sample of gene trees, since they result from methodological artifacts or rare non-vertical events. Exclusion of root candidates, however, entails the reduction of the total number of tested candidates, whereas the use of pseudocounts presents a more inclusive solution for the problem.

The accuracy of our tests is well demonstrated in the cyanobacteria and opisthokonta datasets, those are our positive controls. For both datasets we retrieve consistently the known root partitions, across different samples of gene trees. The opisthokonta root shows the monophyly of fungi and metazoan species. The cyanobacteria root is in agreement with a unicellular LCA, with multicellularity being a trait that evolved later. This is in contrast to reports in the literature of an early-origin of multicellularity within the cyanobacteria phylum. It is notable that those studies were based on cyanobacteria phylogenies rooted with inappropriate outgroups (e.g., (Schirrmeister et al. 2011)).

Our root inference approach resolves an apparent root neighborhood observed in proteobacteria (Tria et al. 2017), and selects the partition that separates epsilonproteobacteria from the other classes as the definitive root partition. The apparent root neighborhood in proteobacteria stems from a tight competition among three candidate root partitions: the epsilonproteobacteria partition, the partition joining epsilon and deltaproteobacteria, and the alphaproteobacteria partition. This apparent root neighborhood is observed when using the majority-rule approach. This approach considers only CSC gene trees and does not use formal tests for the selection of the best root partition(s). The competition among alternative root partitions in datasets with moderate root signal, like the proteobacteria dataset, can be resolved only by comparing the *r* values for all the branches in the gene trees in a statistical framework.

The inferred root partition for the proteobacteria dataset indicates that the proteobacteria LCA is a closer relative of epsilonproteobacteria in comparison to the other classes. Species

belonging to the epsilonproteobacteria class are generally anaerobes and their energy metabolism is based on alternative electron acceptors to oxygen. For example, *Wolinella succinogenes* can perform oxidative phosphorylation with fumarate as terminal electron acceptor, a process known as fumarate respiration (Baar et al. 2003). Another example is the *Sulfurospirillum deleyianum* that can perform anaerobic respiration using various electron acceptors (Sievert et al. 2008). In addition, epsilonproteobacteria species show versatile biochemical strategies to fix carbon, enabling members of this class to colonize extreme environments such as deep sea hydrothermal vents (for review, see (Campbell et al. 2006)). In a previous study, numerous epsilonproteobacteria isolates were retrieved from deep sea hydrothermal vents, most of them characterized as chemolithoautotrophs (Takai et al. 2005). Taken together, these observations, and the epsilonproteobacteria root, suggest that: 1) the proteobacteria LCA was anaerobe and aerobic respiration evolved later in the phylum; 2) the proteobacteria LCA was likely a chemolithoautotrophic lineage inhabiting an extreme environment, with heterotrophic lineages appearing as later innovations. The inference of chemolithoautotrophic and anaerobic life-style for ancient lineages, such as the proteobacteria LCA, is in line with the scenario of life's early phase as predicted by the hydrothermal-vent theory for the origin of life (Martin et al. 2008).

The root neighborhood inferred for the archaea dataset supplies reconciliation for the conflicting reports in the literature (Woese et al. 1990; Waters et al. 2003; Raymann et al. 2015; Williams et al. 2017). Those previous studies used phylogenomic rooting approaches that do not account for inferences of root neighborhoods. Consequently, they do not explicitly embrace methodological uncertainty or evolutionary scenarios incongruent with a bifurcating species tree. The archaea root neighborhood observed here includes root partitions that correspond to previous reports. Raymann *et al.* (2015) placed the root of archaea within the euryarchaeota, a situation congruent with three root partitions present in our root neighborhood: one partition separating halobacteria from the other archaea, one partition joining halobacteria and mathanomicrobia, and one partition separating archaeoglobi, methanococci and thermoplasmata from the other species. Williams *et al.* (2016) reported the branch leading to the DPANN clade as the root branch of the archaea species tree. In our species set, the DPANN clade is represented by the *Nanoarchaeum equitans* species. Our root neighborhood contains a partition that separates the *Nanoarchaeum equitans* species from the others. Besides, the branch leading *Nanorachaeum equitans* had been previously reported as the root of the species tree elsewhere (Waters et al. 2003). In addition, our root neighborhood also includes a partition that separates crenarchaeota (thermoprotei) and *N. equitans* from the rest of the archaea, a root partition in agreement with one of the pioneering studies regarding the archaea root (Woese et al. 1990). We speculate that the root neighborhood in the archaea dataset stems from reticulated evolution, with genes tracing back to a plurality of ancestor genomes (LCA). Indeed, previous studies suggested massive LGT from bacteria to the

ancestors of the major archaea phyla (Nelson-Sathi et al. 2012; Nelson-Sathi et al. 2015). Massive LGT has the potential of creating chimeric genomes with sets of genes tracing back to different LCAs as we discussed here. In summary, our results supply further evidence that modern archaea genomes descended from multiple LCAs.

5.4 Methodology

Protein families for the opisthokonta, proteobacteria and archaea datasets were extracted from EggNOG version 4.5 (Huerta-Cepas et al. 2016). The cyanobacteria protein families were constructed from completely sequenced genomes available in RefSeq database (O'Leary et al. 2016) (see section 4.1 for detailed information). The datasets are: Opisthokonta (31 species), Proteobacteria (72 species), Cyanobateria (130 species) and Archaea (115 species) (See Supplementary Table 3 for the complete list of species). For each of the 4 datasets we retrieved all the protein families present in at least three species. In this study we were interested in multi-copy and single-copy protein families, as well as partial and complete protein families. The distribution of protein family category in the 4 datasets can be found in Table 4.

Protein sequences of the resulting protein families were aligned using MAFFT ver. v7.027b with L-INS-i alignment strategy (Katoh and Standley 2013). Phylogenetic trees were reconstructed using PhyML version 20120412 (Guindon et al. 2010) with the following parameters: -b -4 -v e -m LG -c 4 -s SPR. MAD rooting was performed using in-house MatLab© scripts.

6 Outlook

The rooting approaches presented here open the way for novel methodological developments and applications to address diverse evolutionary questions.

From a methodological perspective, the MAD rooting approach holds promise for the improvement of tree reconstruction algorithms. A possible research direction is to combine MAD rooting with existing tree reconstruction methods. The goal would be to devise a novel tree reconstruction procedure that yields a rooted tree topology as an output. We suggest a bottom-up approach, where the tree reconstruction problem is seen as a series of rooting problems. Given an OTU set, the first root to be in inferred is the global root of the tree. The inference of the global root determines two clades; those clades are the OTUs on either side of the global root. Each resulting clade is, again, input for the root inference. This process can continue in a recursive manner, through all local roots of the phylogenetic tree. The recursive rooting of phylogenies may improve the quality of reconstructed trees, because in each step of the recursion only the relevant OTUs are considered for root inferences. Methodological artifacts from common phylogenetic tree reconstruction methods become more acute in datasets with many OTUs, due to drastic increase in computational complexity of finding the best phylogenetic tree. Interestingly, the rooted tree reconstruction we propose is agnostic with respect to the framework of choice (i.e., neighbor-joining, maximum parsimony, maximum likelihood or Bayesian), evolutionary model or data type. The bottom-up rooted tree reconstruction is applicable for single gene trees and also in a phylogenomic context, when the goal to infer a species tree. The inference of the species tree may be done by using the total evidence from gene trees to recursively infer root partitions as discussed in section 5.

From the biological perspective, the *r* statistic from MAD enables the comparison of alternative ancestor-descendent relations among ancestral lineages. One example of practical application is, for instance, to establish the chronological order of trait appearance in a species set. For a pair of traits, each tracing back to its LCA, one can evaluate their relative chronological priority from a sample of gene trees. The evaluation of alternative ancestor-descendent relations between LCAs is possible by comparing the *r* statistics for different root positions in the individual gene trees. This approach can be applied to address evolutionary hypotheses without requiring a species tree (for example, questions similar to the marine-freshwater origin of cyanobacteria discussed in section 4). We expect that the quantification of alternative ancestor-descendent relations with the *r* statistics will also allow the inference of directed LGT, gene duplications and the reconstruction of biochemical pathways for ancestor genomes.

Our studies enabled the identification of root neighborhoods, at both the gene tree and species tree level. Accounting for root neighborhoods paves the way for new discoveries in

evolution, outside the constraints of a single, 'hard coded', LCA. The notion of single roots results from the assumption of an exclusively bifurcating process that underlies the evolution of biological entities. If evolution is, in fact, non-bifurcating a single root will inevitably show an incorrect picture of ancestor-descendent relations. For instance, LGT is prevalent in prokaryotes, a notable violation of the assumed bifurcating process. Hence, gene trees may differ in topology and roots among themselves and a species trees (e.g., (Bapteste et al. 2009)). Here we embrace the plurality of LCAs, accounting for non-bifurcating processes underlying the evolution of biological entities. In this regard, our studies conform to the 'pattern pluralism' doctrine, forcefully elaborated in (Doolittle and Bapteste 2007).

In the RPE procedure the plurality of LCAs is the null hypothesis that we test. If the molecular data does not support the existence of multiple LCAs, then the alternative hypothesis of a single LCA is preferred. It is remarkable that we observe examples of prokaryotic taxa with single LCAs and one example with multiple LCAs. The taxon with multiple LCAs demonstrates that a completely bifurcating tree of life for prokaryotes does not exist. The taxa with single LCAs, however, pinpoint instances of bifurcating processes along prokaryotic evolution.

7 References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Baar C, Eppinger M, Raddatz G, Simon J, Lanz C, Klimmek O, Nandakumar R, Gross R, Rosinus A, Keller H, et al. 2003. Complete genome sequence and analysis of *Wolinella succinogenes*. Proc. Natl. Acad. Sci. U.S.A. 100:11690–11695.

Bapteste E, O'Malley MA, Beiko RG, Ereshefsky M, Gogarten JP, Franklin-Hall L, Lapointe F-J, Dupré J, Dagan T, Boucher Y, et al. 2009. Prokaryotic evolution and the tree of life are two different things. Biology Direct 4:34.

Campbell BJ, Engel AS, Porter ML, Takai K. 2006. The versatile epsilon-proteobacteria: key players in sulphidic habitats. Nat. Rev. Mol. Cell Biol. 4:458–468.

Dagan T, Artzy-Randrup Y, Martin W. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc. Natl. Acad. Sci. U.S.A. 105:10039–10044.

Dagan T, Roettger M, Stucken K, Landan G, Koch R, Major P, Gould SB, Goremykin VV, Rippka R, Tandeau de Marsac N, et al. 2013. Genomes of Stigonematalean Cyanobacteria (Subsection V) and the Evolution of Oxygenic Photosynthesis from Prokaryotes to Plastids. Genome Biol. Evol. 5:31–44.

Doolittle WF, Bapteste E. 2007. Pattern pluralism and the Tree of Life hypothesis. Proc. Natl. Acad. Sci. U.S.A. 104:2043–2049.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nuc. Acids Res. 30:1575–1584.

Farris JS. 1972. Estimating Phylogenetic Trees from Distance Matrices. The American Naturalist 106:645–668.

Fitch WM, Margoliash E. 1967. Construction of phylogenetic trees. 155:279–284.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biol. 59:307–321.

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, et al. 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nuc. Acids Res. 44:D286–D293.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772–780.

Katz LA, Grant JR, Parfrey LW, Burleigh JG. 2012. Turning the crown upside down: gene tree parsimony roots the eukaryotic tree of life. Systematic Biol. 61:653–660.

Kluge AG, Farris JS. 1969. Quantitative Phyletics and the Evolution of Anurans. Systematic Biol. 18:1–32.

Ku C, Nelson-Sathi S, Roettger M, Garg S, Hazkani-Covo E, Martin WF. 2015. Endosymbiotic gene transfer from prokaryotic pangenomes: Inherited chimerism in eukaryotes. Proc. Natl. Acad. Sci. U.S.A.:201421385.

Lepage T, Bryant D, Philippe H, Lartillot N. 2007. A general comparison of relaxed molecular clock models. Mol. Biol. Evol. 24:2669–2680.

Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Pillay M, Ratner A, Huang J, Woyke T, Huntemann M, et al. 2014. IMG 4 version of the integrated microbial genomes comparative analysis system. Nuc. Acids Res. 42:D560–D567.

Martin W, Baross J, Kelley D, Russell MJ. 2008. Hydrothermal vents and the origin of life. Nat. Rev. Microbiol. 6:805–814.

Martin WF, Garg S, Zimorski V. 2015. Endosymbiotic theories for eukaryote origin. Philos. Trans. R. Soc. B Biol. Sci. 370:20140330.

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. 2005. The microbial pan-genome. Current Opinion in Genetics & Development 15:589–594.

Nelson-Sathi S, Dagan T, Landan G, Janssen A, Steel M, McInerney JO, Deppenmeier U, Martin WF. 2012. Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. Proc. Natl. Acad. Sci. U.S.A. 109:20537–20542.

Nelson-Sathi S, Sousa FL, Roettger M, Lozada-Chávez N, Thiergart T, Janssen A, Bryant D, Landan G, Schönheit P, Siebers B, et al. 2015. Origins of major archaeal clades correspond to gene acquisitions from bacteria. Nature 517:77–80.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nuc. Acids Res. 44:D733–D745.

Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T. 2011. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. Genome Res. 21:599–609.

Puigbò P, Wolf YI, Koonin EV. 2009. Search for a “Tree of Life” in the thicket of the phylogenetic forest. J Biol 8:59.

Ragan MA. 2009. Trees and networks before and after Darwin. Biology Direct 4:43.

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

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends in Genetics 16:276–277.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biol. 61:539–542.

Schirrmeister BE, Antonelli A, Bagheri HC. 2011. The origin of multicellularity in cyanobacteria. BMC Evol. Biol. 11:45.

Semple C, Steel M. 2000. Tree Reconstruction via a Closure Operation on Partial Splits. In: Computational Biology. Springer, Berlin, Heidelberg. pp. 126–134.

Shih PM, Wu D, Latifi A, Axen SD, Fewer DP, Talla E, Calteau A, Cai F, Tandeau de Marsac N, Rippka R, et al. 2013. Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. Proc. Natl. Acad. Sci. U.S.A. 110:1053–1058.

Sievert SM, Scott KM, Klotz MG, Chain PSG, Hauser LJ, Hemp J, Hügler M, Land M, Lapidus A, Larimer FW, et al. 2008. Genome of the epsilonproteobacterial chemolithoautotroph *Sulfurimonas denitrificans*. Applied Environ. Microbiol. 74:1145–1156.

Stechmann A, Cavalier-Smith T. 2002. Rooting the eukaryote tree by using a derived gene fusion. Science 297:89–91.

Swenson KM, El-Mabrouk N. 2012. Gene trees and species trees: irreconcilable differences. BMC Bioinformatics 13 Suppl 19:S15.

Szöllosi GJ, Tannier E, Daubin V, Boussau B. 2015. The inference of gene trees with species trees. Systematic Biol. 64:e42–e62.

Takai K, Campbell BJ, Cary SC, Suzuki M, Oida H, Nunoura T, Hirayama H, Nakagawa S, Suzuki Y, Inagaki F, et al. 2005. Enzymatic and genetic characterization of carbon and energy metabolisms by deep-sea hydrothermal chemolithoautotrophic isolates of *Epsilonproteobacteria*. Applied Environ. Microbiol. 71:7310–7320.

Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science 278:631–637.

Thiergart T, Landan G, Martin WF. 2014. Concatenated alignments and the case of the disappearing tree. BMC Evolutionary Biology 14:2624.

Tria FDK, Landan G, Dagan T. 2017. Phylogenetic rooting using minimal ancestor deviation. Nat.

Ecol. Evol. 1:0193–0197.

Turner S, Pryer KM, Miao VP, Palmer JD. 1999. Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. J. Eukaryot. Microbiol. 46:327–338.

Waters E, Hohn MJ, Ahel I, Graham DE, Adams MD, Barnstead M, Beeson KY, Bibbs L, Bolanos R, Keller M, et al. 2003. The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism. Proc. Natl. Acad. Sci. U.S.A. 100:12984–12988.

West SA, Fisher RM, Gardner A, Kiers ET. 2015. Major evolutionary transitions in individuality. Proc. Natl. Acad. Sci. U.S.A. 112:10112–10119.

Williams TA, Heaps SE, Cherlin S, Nye TMW, Boys RJ, Embley TM. 2015. New substitution models for rooting phylogenetic trees. Philos. Trans. R. Soc. B Biol. Sci. 370:20140336.

Williams TA, Szöllosi GJ, Spang A, Foster PG, Heaps SE, Boussau B, Ettema TJG, Embley TM. 2017. Integrative modeling of gene and genome evolution roots the archaeal tree of life. Proc. Natl. Acad. Sci. U.S.A. 114:E4602–E4611.

Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. U.S.A. 87:4576–4579.

8 Acknowledgments

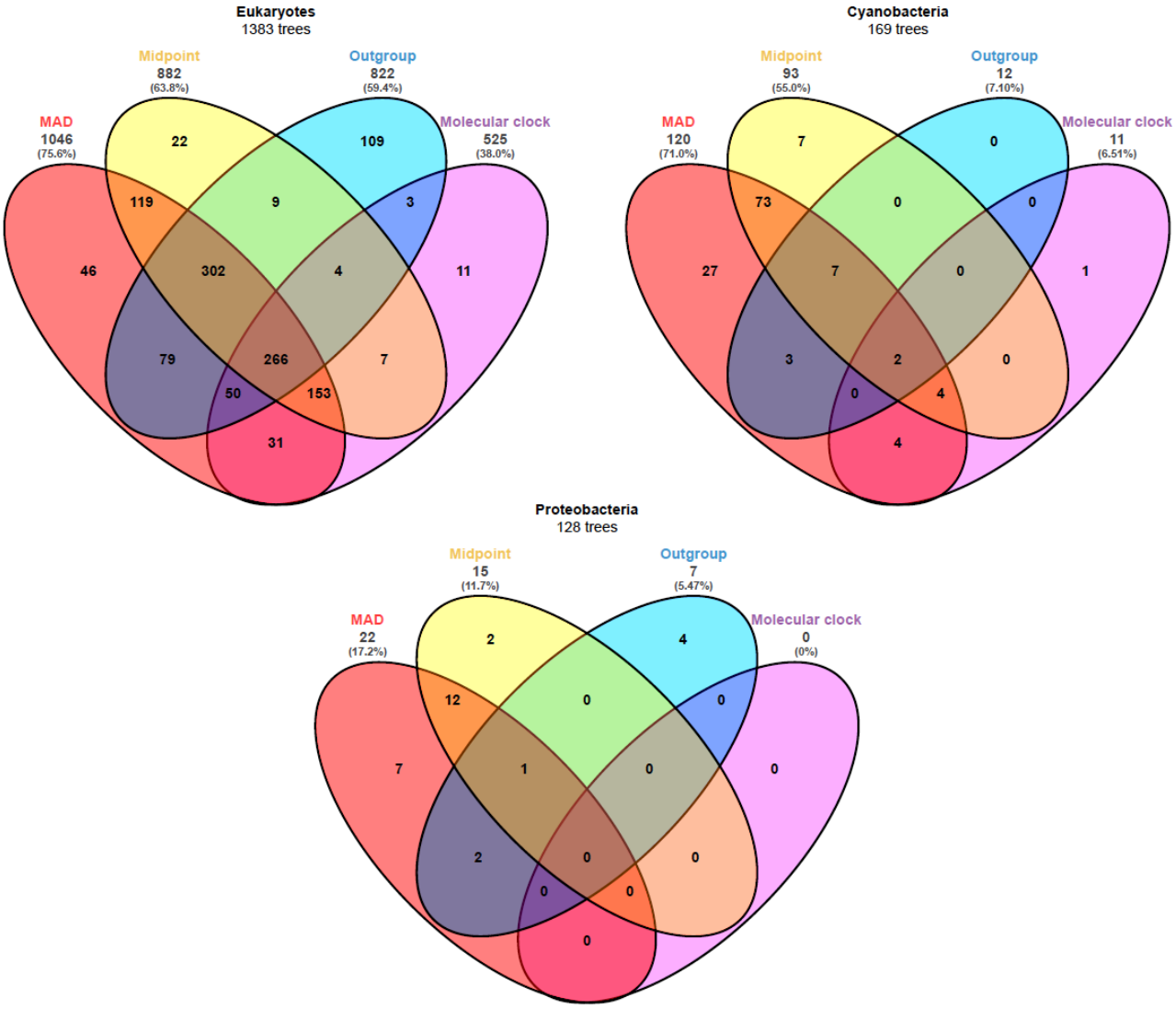
The accomplishment of this thesis wouldn't be possible without the steadfast support from my family: Hozana, Paulo and Tiago. Thank you!

I reserve a special thanks to my mentors Tal Dagan and Giddy Landan for all the teaching and guidance. I also got valuable help from all the members of the research group in innumerous situations (in science and in life).

My friends, some geographically far, were crucial to give me the necessary balance to go through this journey.

I would like to acknowledge the institutions CAPES (Coordination of Improvement of Higher Education Personnel - Brazil) and ERC (European Research Council) for the financial support of the studies presented here.

9 Supplementary



Supplementary Figure 1: Cross performance of four rooting methods in three datasets. Only the best root branch in each dataset is presented. The set of genes consists of all genes where the split is present in at least one of the three tree reconstructions (MAD and midpoint are based on the same ML tree).

Supplementary Table 1: Rooting performance of MAD variants, molecular clock, midpoint and outgroup for the three datasets. The ten most frequently inferred root branches are presented as serial numbers, following the order of branches presented in Figure 3. Values are percentage of trees rooted on each of the branches, the proportion of unrootable trees or the percentage of alternative root branches (other). The percentage of ML trees containing the branches is indicated (branch frequency).

a. Eukaryotes (n=1446)	Branch											
Method	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Unrootable**	**Other**
MAD=A1-B1	72.34	10.37	1.73	1.11	2.07	0.14	0.14	0.00	0.14	0.00	0.00	11.96
A1-B2	55.67	13.42	5.74	2.35	6.78	0.35	0.48	0.00	0.14	0.07	0.00	15.01
A2-B1	68.67	12.45	1.66	0.76	3.04	0.14	0.14	0.00	0.14	0.00	0.00	13.00
A2-B2	55.39	14.18	5.12	2.14	6.98	0.28	0.41	0.00	0.14	0.07	0.00	15.28
MCCV	62.10	14.52	1.52	0.48	6.09	0.14	0.14	0.07	0.21	0.00	0.41	14.32
PMR	58.51	16.18	1.52	0.48	8.23	0.14	0.14	0.07	0.21	0.00	0.07	14.45
Midpoint	61.00	12.10	2.21	1.87	4.56	0.14	0.21	0.00	0.07	0.07	0.00	17.77
Molecular clock	36.31	9.96	1.66	34.30	0.00	0.07	0.14	0.07	0.07	0.00	0.00	17.43
Outgroup	56.85	1.31	0.90	3.73	3.80	1.04	0.69	0.00	0.00	0.48	20.82	10.37
Branch frequency	94.40	97.17	89.90	50.35	100.0	51.31	85.75	36.58	100.0	8.64		

b. Cyanobacteria (n=172)	Branch											
Method	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Unrootable**	**Other**
MAD=A1-B1	69.77	8.72	1.74	0.58	0.00	0.58	0.00	2.33	0.00	0.00	0.00	16.28
A1-B2	65.70	6.40	1.16	0.58	0.58	2.33	0.58	2.91	1.16	0.00	0.00	18.60
A2-B1	69.19	9.30	1.16	1.16	0.58	1.74	0.00	2.33	0.00	0.00	0.00	14.54
A2-B2	62.79	6.98	1.16	1.16	2.33	2.33	0.00	3.49	1.16	0.58	0.00	18.02
MCCV	43.02	0.58	0.00	0.00	0.00	2.33	0.00	0.00	0.58	0.58	27.33	25.58
PMR	59.88	4.07	0.00	0.00	1.16	2.91	0.58	12.21	1.74	0.58	0.00	16.87
Midpoint	54.07	6.98	2.33	0.58	2.33	1.16	0.00	0.00	0.58	0.00	0.00	31.98
Molecular clock	6.40	0.00	1.16	14.54	0.58	0.00	0.00	0.00	0.00	0.00	27.91	49.42
Outgroup:												
G. violaceus	6.98	1.16	0.00	0.00	0.00	0.00	5.23	37.79	6.40	4.65	4.65	33.14
G. kilaueensis	8.14	1.74	0.00	0.00	0.00	0.00	6.40	33.72	6.40	6.40	11.05	26.16
M. Zag 1	10.47	1.16	0.00	0.00	0.00	2.91	3.49	11.63	1.16	3.49	43.02	22.67
L. buccalis	4.65	2.33	0.00	0.00	0.00	2.33	1.74	5.81	2.91	5.81	47.67	26.74
C. aurantiacus	11.05	1.16	0.00	0.00	0.00	0.58	1.16	6.40	4.65	8.14	33.72	33.14
E. coli	13.37	0.00	0.00	0.00	0.58	1.74	1.16	6.40	2.91	5.81	41.86	26.16
Branch frequency	98.26	50.00	25.58	33.72	79.65	97.67	81.98	81.40	100.0	97.67		

c. Proteobacteria (n=130)	Branch											
Method	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Unrootable**	**Other**
MAD=A1-B1	16.92	1.54	8.46	4.62	13.85	2.31	1.54	0.77	1.54	1.54	0.00	46.92
A1-B2	24.62	1.54	3.08	5.38	9.23	3.85	3.08	3.08	2.31	0.77	0.00	43.08
A2-B1	21.54	1.54	7.69	3.08	10.00	2.31	2.31	1.54	0.77	0.77	0.00	48.46
A2-B2	22.31	1.54	3.85	4.62	8.46	2.31	3.08	3.08	2.31	0.77	0.00	47.69
MCCV	26.92	0.00	3.08	1.54	8.46	1.54	2.31	2.31	0.00	0.00	17.69	36.15
PMR	32.31	0.00	1.54	6.15	10.77	3.08	3.08	3.08	2.31	0.00	0.00	37.69
Midpoint	11.54	3.85	10.00	2.31	6.15	0.77	2.31	3.08	2.31	1.54	0.00	56.15
Molecular clock	0.00	4.62	11.54	1.54	3.08	0.00	0.00	0.00	0.00	0.77	3.08	75.39
Outgroup	5.38	0.77	3.08	0.00	0.77	0.00	0.00	0.00	0.00	0.77	73.85	15.38
Branch frequency	98.46	26.15	40.00	67.69	84.62	90.77	77.69	100.0	12.31	8.46		

Supplementary Table 2: Rooting performance of MAD, molecular clock, midpoint and outgroup; for gene families with outgroup homologs. The ten most frequently inferred root branches are presented as serial numbers, following the order of branches presented in Figure 3. Values are percentage of trees rooted on each of the branches, the proportion of unrootable trees or the percentage of alternative root branches (other).

a. Eukaryotes (n=1249)	Branch											
Method	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Unrootable**	**Other**
MAD	71.10	11.05	1.84	0.88	2.16	0.16	0.16	0.00	0.16	0.00	0.00	12.49
Midpoint	59.09	12.57	2.48	1.84	4.88	0.16	0.24	0.00	0.08	0.08	0.00	18.58
Molecular clock	34.99	10.01	1.84	34.59	0.00	0.08	0.16	0.00	0.08	0.00	0.00	18.26
Outgroup	65.81	1.52	1.04	4.32	4.40	1.20	0.80	0.00	0.00	0.56	8.33	12.01

b. Cyanobacteria (n=67)	Branch											
Method	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Unrootable**	**Other**
MAD	62.69	16.42	1.49	0.00	0.00	1.49	0.00	1.49	0.00	0.00	0.00	16.42
Midpoint	47.76	10.45	4.48	1.49	1.49	1.49	0.00	0.00	1.49	0.00	0.00	31.34
Molecular clock	0.00	0.00	0.00	13.43	1.49	0.00	0.00	0.00	0.00	0.00	25.37	59.70
Outgroup:												
G. violaceus	4.48	1.49	0.00	0.00	0.00	0.00	2.99	44.78	5.97	7.46	0.00	32.84
G. kilaueensis	5.97	4.48	0.00	0.00	0.00	0.00	5.97	40.30	5.97	8.96	0.00	28.36
M. Zag 1	22.39	2.99	0.00	0.00	0.00	4.48	2.99	19.40	2.99	7.46	0.00	37.31
L. buccalis	10.45	5.97	0.00	0.00	0.00	4.48	2.99	14.93	4.48	8.96	0.00	47.76
C. aurantiacus	13.43	2.99	0.00	0.00	0.00	1.49	2.99	7.46	7.46	13.43	0.00	50.75
E. coli	20.90	0.00	0.00	0.00	0.00	4.48	0.00	14.93	4.48	10.45	0.00	44.78

c. Proteobacteria (n=102)	Branch											
Method	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Unrootable**	**Other**
MAD	17.65	1.96	8.82	3.92	15.69	2.94	0.98	0.98	1.96	1.96	0.00	43.14
Midpoint	13.73	3.92	10.78	0.98	6.86	0.98	1.96	1.96	3.92	1.96	0.00	52.94
Molecular clock	0.00	3.92	13.73	1.96	3.92	0.00	0.00	0.00	0.00	0.98	2.94	72.55
Outgroup	6.86	0.98	3.92	0.00	0.98	0.00	0.00	0.00	0.00	0.98	66.67	19.61

Supplementary Table 3: Species composition in the four datasets. For eukaryotes (opisthokonta), proteobacteria and cyanobacteria the ingroup species are ordered according to their corresponding position in the black and white matrix of Figure 3. EggNOG identifiers are provided for the proteobacteria, archaea and eukaryotic datasets. For the cyanobacteria dataset the NCBI assembly accessions are provided, except for Melainabacteria Zag 1 for which the IMG genome identifier is provided.

a	**Eukaryotes (Opisthokonta)**	
Position	**Name**	**EggNogID**
1	Caenorhabditis elegans	6239
2	Nasonia vitripennis	7425
3	Apis mellifera	7460
4	Pediculus humanus	121225
5	Drosophila melanogaster	7227
6	Anopheles gambiae	7165
7	Strongylocentrotus purpuratus	7668
8	Oreochromis niloticus	8128
9	Oryzias latipes	8090
10	Danio rerio	7955
11	Xenopus (Silurana) tropicalis	8364
12	Mus musculus	10090
13	Homo sapiens	9606
14	Otolemur garnettii	30611
15	Podospora anserina S mat	515849
16	Neurospora crassa	5141
17	Sordaria macrospora k hell	771870
18	Magnaporthe oryzae	318829
19	Fusarium oxysporum	5507
20	Phaeosphaeria nodorum	13684
21	Penicillium chrysogenum Wisconsin 54 1255	500485
22	Aspergillus clavatus	5057
23	Talaromyces stipitatus ATCC 10500	441959
24	Ajellomyces dermatitidis SLH14081	559298
25	Coccidioides posadasii C735 delta SOWgp	222929
26	Trichophyton rubrum CBS 118892	559305
27	Arthroderma gypseum CBS 118893	535722
28	Lachancea thermotolerans	381046
29	Kluyveromyces lactis NRRL Y 1140	284590
30	Torulaspora delbrueckii	4950
31	Candida glabrata CBS 138	284593
	Outgroup: Selaginella moellendorffii	88036
	Outgroup: Physcomitrella patens	3218
	Outgroup: Brachypodium distachyon	15368
	Outgroup: Oryza sativa Japonica Group	39947
	Outgroup: Zea mays	4577
	Outgroup: Solanum lycopersicum	4081
	Outgroup: Vitis vinifera	29760
	Outgroup: Arabidopsis thaliana	3702
	Outgroup: Populus trichocarpa	3694
	Outgroup: Glycine max	3847

b	**Cyanobacteria**	

Position	Name	Assemby accession
1	Mastigocladopsis repens PCC 10914	GCA_000315565.1
2	Fischerella sp. PCC 9605	GCA_000517105.1
3	Fischerella muscicola PCC 7414	GCA_000317205.1
4	Fischerella thermalis PCC 7521	GCA_000317225.1
5	Fischerella sp. JSC-11	GCA_000231365.2
6	Fischerella sp. PCC 9431	GCA_000447295.1
7	Fischerella sp. PCC 9339	GCA_000315585.1
8	Fischerella muscicola SAG 1427-1 = PCC 73103	GCA_000317245.1
9	cyanobacterium PCC 7702	GCA_000332255.1
10	Chlorogloeopsis fritschii PCC 6912	GCA_000317285.1
11	Chlorogloeopsis fritschii PCC 9212	GCA_000317265.1
12	Scytonema hofmanni UTEX 2349	GCA_000582685.1
13	Nostoc sp. PCC 7120	GCA_000009705.1
14	Anabaena variabilis ATCC 29413	GCA_000204075.1
15	Nostoc sp. PCC 7524	GCA_000316645.1
16	Nostoc sp. PCC 7107	GCA_000316625.1
17	Nodularia spumigena CCY9414	GCA_001586755.1
18	Nostoc punctiforme PCC 73102	GCA_000020025.1
19	Cylindrospermum stagnale PCC 7417	GCA_000317535.1
20	Nostoc azollae 0708	GCA_000196515.1
21	Cylindrospermopsis raciborskii CS-505	GCA_000175835.1
22	Raphidiopsis brookii D9	GCA_000175855.1
23	Anabaena sp. PCC 7108	GCA_000332135.1
24	Anabaena cylindrica PCC 7122	GCA_000317695.1
25	Calothrix sp. PCC 7507	GCA_000316575.1
26	Microchaete sp. PCC 7126	GCA_000332295.1
27	Rivularia sp. PCC 7116	GCA_000316665.1
28	Calothrix sp. PCC 7103	GCA_000331305.1
29	Calothrix sp. PCC 6303	GCA_000317435.1
30	Scytonema hofmanni PCC 7110	GCA_000346485.2
31	Pseudanabaena sp. PCC 7367	GCA_000317065.1
32	Pseudanabaena biceps PCC 7429	GCA_000332215.1
33	Pseudanabaena sp. PCC 6802	GCA_000332175.1
34	Geitlerinema sp. PCC 7407	GCA_000317045.1
35	Leptolyngbya sp. JSC-1	GCA_000733415.1
36	Leptolyngbya sp. PCC 6406	GCA_000332095.2
37	Nodosilinea nodulosa PCC 7104	GCA_000309385.1
38	Leptolyngbya sp. PCC 7375	GCA_000316115.1
39	Leptolyngbya boryana PCC 6306	GCA_000353285.1
40	Oscillatoriales cyanobacterium JSC-12	GCA_000309945.1
41	Geitlerinema sp. PCC 7105	GCA_000332355.1
42	Oscillatoria acuminata PCC 6304	GCA_000317105.1
43	Oscillatoria sp. PCC 10802	GCA_000332335.1
44	Oscillatoria nigro-viridis PCC 7112	GCA_000317475.1
45	Oscillatoria sp. PCC 6506	GCA_000180455.1
46	Kamptonema formosum PCC 6407	GCA_000332155.1
47	Trichodesmium erythraeum IMS101	GCA_000014265.1
48	Arthrospira platensis NIES-39	GCA_000210375.1
49	Arthrospira maxima CS-328	GCA_000173555.1
50	Arthrospira sp. PCC 8005	GCA_000176895.2
51	Arthrospira platensis C1	GCA_000307915.1

52	Lyngbya sp. PCC 8106	GCA_000169095.1
53	Crinalium epipsammum PCC 9333	GCA_000317495.1
54	Coleofasciculus chthonoplastes PCC 7420	GCA_000155555.1
55	Microcoleus sp. PCC 7113	GCA_000317515.1
56	filamentous cyanobacterium ESFC-1	GCA_000380225.1
57	Spirulina subsalsa PCC 9445	GCA_000314005.1
58	Leptolyngbya sp. PCC 7376	GCA_000316605.1
59	Prochlorothrix hollandica PCC 9006	GCA_000341585.2
60	Chroococcidiopsis thermalis PCC 7203	GCA_000317125.1
61	Pleurocapsa sp. PCC 7327	GCA_000317025.1
62	Stanieria cyanosphaera PCC 7437	GCA_000317575.1
63	Xenococcus sp. PCC 7305	GCA_000332055.1
64	Pleurocapsa sp. PCC 7319	GCA_000332195.1
65	Synechococcus sp. RCC307	GCA_000063525.1
66	Synechococcus sp. WH 5701	GCA_000153045.1
67	Cyanobium sp. PCC 7001	GCA_000155635.1
68	Cyanobium gracile PCC 6307	GCA_000316515.1
69	Synechococcus sp. CB0101	GCA_000179235.1
70	Synechococcus sp. CB0205	GCA_000179255.1
71	Synechococcus sp. CC9616	GCA_000515235.1
72	Synechococcus sp. WH 8102	GCA_000195975.1
73	Synechococcus sp. WH 8109	GCA_000161795.2
74	Synechococcus sp. CC9605	GCA_000012625.1
75	Synechococcus sp. BL107	GCA_000153805.1
76	Synechococcus sp. CC9902	GCA_000012505.1
77	Synechococcus sp. RS9916	GCA_000153825.1
78	Synechococcus sp. RS9917	GCA_000153065.1
79	Synechococcus sp. WH 8016	GCA_000230675.2
80	Synechococcus sp. CC9311	GCA_000014585.1
81	Synechococcus sp. WH 7803	GCA_000063505.1
82	Synechococcus sp. WH 7805	GCA_000153285.1
83	Prochlorococcus marinus str. MIT 9313	GCA_000011485.1
84	Prochlorococcus marinus str. MIT 9303	GCA_000015705.1
85	Prochlorococcus marinus str. MIT 9211	GCA_000018585.1
86	Prochlorococcus marinus subsp. marinus str. CCMP1375	GCA_000007925.1
87	Prochlorococcus marinus str. NATL2A	GCA_000012465.1
88	Prochlorococcus marinus str. NATL1A	GCA_000015685.1
89	Prochlorococcus marinus str. MIT 9215	GCA_000018065.1
90	Prochlorococcus marinus str. MIT 9202	GCA_000158595.1
91	Prochlorococcus marinus str. AS9601	GCA_000015645.1
92	Prochlorococcus marinus str. MIT 9301	GCA_000015965.1
93	Prochlorococcus marinus str. MIT 9312	GCA_000012645.1
94	Prochlorococcus marinus str. MIT 9515	GCA_000015665.1
95	Prochlorococcus marinus subsp. pastoris	GCA_000011465.1
96	Synechococcus sp. JA-3-3Ab	GCA_000013205.1
97	Synechococcus sp. JA-2-3B'a(2-13)	GCA_000013225.1
98	Synechococcus sp. PCC 7336	GCA_000332275.1
99	Acaryochloris marina MBIC11017	GCA_000018105.1
100	Acaryochloris sp. CCMEE 5410	GCA_000238775.2
101	Cyanothece sp. PCC 7425	GCA_000022045.1
102	Thermosynechococcus elongatus BP-1	GCA_000011345.1
103	Synechococcus sp. PCC 6312	GCA_000316685.1

104	Synechococcus sp. PCC 7335	GCA_000155595.1
105	Gloeocapsa sp. PCC 7428	GCA_000317555.1
106	Synechocystis sp. PCC 7509	GCA_000332075.2
107	Chamaesiphon minutus PCC 6605	GCA_000317145.1
108	Dactylococcopsis salina PCC 8305	GCA_000317615.1
109	Halothece sp. PCC 7418	GCA_000317635.1
110	Gloeocapsa sp. PCC 73106	GCA_000332035.1
111	Microcystis aeruginosa NIES-843	GCA_000010625.1
112	Cyanothece sp. PCC 7822	GCA_000147335.1
113	Cyanothece sp. PCC 7424	GCA_000021825.1
114	Cyanothece sp. PCC 8802	GCA_000024045.1
115	Cyanothece sp. PCC 8801	GCA_000021805.1
116	Candidatus Atelocyanobacterium thalassa	GCA_000025125.1
117	Crocosphaera watsonii WH 8501	GCA_000167195.1
118	Cyanothece sp. CCY0110	GCA_000169335.1
119	Cyanothece sp. ATCC 51142	GCA_000017845.1
120	Cyanothece sp. ATCC 51472	GCA_000231425.3
121	Synechocystis sp. PCC 6803 substr. GT-I	GCA_000284135.1
122	Synechocystis sp. PCC 6803 substr. PCC-N	GCA_000284215.1
123	Synechocystis sp. PCC 6803	GCA_001318385.1
124	Synechocystis sp. PCC 6803 GT-S	GCA_000270265.1
125	Cyanobacterium stanieri PCC 7202	GCA_000317655.1
126	Cyanobacterium aponinum PCC 10605	GCA_000317675.1
127	Geminocystis herdmanii PCC 6308	GCA_000332235.1
128	Synechococcus sp. PCC 7002	GCA_000019485.1
129	Synechococcus elongatus PCC 7942	GCA_000012525.1
130	Synechococcus elongatus PCC 6301	GCA_000010065.1
	Outgroup: Gloeobacter violaceus PCC 7421	GCA_000011385.1
	Outgroup: Gloeobacter kilaueensis	GCA_000484535.1
	Outgroup: Melainabacteria Zag 1	2523533517
	Outgroup: Leptotrichia buccalis	GCA_000023905.1
	Outgroup: Chloroflexus aurantiacus J-10-fl	GCA_000018865.1
	Outgroup: Escherichia coli str. K-12 substr. MG1655	GCA_000005845.2

c Position	Proteobacteria Name	EggNogID
1	Sulfurospirillum deleyianum DSM 6946	525898
2	Sulfurimonas denitrificans DSM 1251	326298
3	Wolinella succinogenes DSM 1740	273121
4	Desulfarculus baarsii DSM 2075	644282
5	Desulfurivibrio alkaliphilus AHT2	589865
6	Desulfobacca acetoxidans DSM 11109	880072
7	Syntrophobacter fumaroxidans MPOB	335543
8	Desulfobacterium autotrophicum HRM2	177437
9	Desulfococcus oleovorans Hxd3	96561
10	Desulfatibacillum alkenivorans AK-01	439235
11	Syntrophus aciditrophicus SB	56780
12	Pelobacter carbinolicus DSM 2380	338963
13	Pelobacter propionicus DSM 2379	338966
14	Desulfovibrio aespoeensis Aspo-2	643562
15	Desulfohalobium retbaense DSM 5692	485915

16	Desulfomicrobium baculatum DSM 4028	525897
17	Desulfovibrio magneticus RS-1	573370
18	Desulfovibrio vulgaris str. 'Miyazaki F'	883
19	Bdellovibrio bacteriovorus HD100	264462
20	Anaeromyxobacter dehalogenans 2CP-C	290397
21	Myxococcus xanthus DK 1622	246197
22	Haliangium ochraceum DSM 14365	502025
23	Sorangium cellulosum 'So ce 56'	448385
24	Acidithiobacillus caldus ATCC 51756	637389
25	Halothiobacillus neapolitanus c2	555778
26	Thioalkalimicrobium cyclicum ALM1	717773
27	Xanthomonas campestris pv. vesicatoria str. 85-10	316273
28	Legionella pneumophila str. Paris	297246
29	Escherichia coli str. K-12 substr. MG1655	511145
30	Acinetobacter baumannii ATCC 19606	575584
31	Pseudomonas putida S16	1042876
32	Hahella chejuensis KCTC 2396	349521
33	gamma proteobacterium IMCC3088	876044
34	Marinomonas posidonica IVIA-Po-181	491952
35	Alkalilimnicola ehrlichii MLHE-1	187272
36	Halorhodospira halophila SL1	349124
37	Allochromatium vinosum DSM 180	572477
38	Methylomonas methanica MC09	857087
39	Methylococcus capsulatus str. Bath	243233
40	Nitrosococcus halophilus Nc4	472759
41	Neisseria meningitidis MC58	122586
42	Methylotenera mobilis JLW8	583345
43	Nitrosospira multiformis ATCC 25196	323848
44	Nitrosomonas eutropha C91	335283
45	Candidatus Accumulibacter phosphatis clade IIA str. UW-1	522306
46	Burkholderia rhizoxinica HKI 454	882378
47	Ralstonia eutropha H16	381666
48	Collimonas fungivorans Ter331	1005048
49	Bordetella petrii DSM 12804	340100
50	Methylibium petroleiphilum PM1	420662
51	Variovorax paradoxus EPS	595537
52	Erythrobacter litoralis HTCC2594	314225
53	Sphingomonas wittichii RW1	392499
54	Sphingobium japonicum UT26S	452662
55	Ruegeria pomeroyi DSS-3	246200
56	Roseobacter litoralis Och 149	391595
57	Parvularcula bermudensis HTCC2503	314260
58	Maricaulis maris MCS10	394221
59	Hyphomonas neptunium ATCC 15444	228405
60	Hirschia baltica ATCC 49814	582402
61	Asticcacaulis excentricus CB 48	573065
62	Brevundimonas subvibrioides ATCC 15264	633149
63	Phenylobacterium zucineum HLK1	450851
64	Parvibaculum lavamentivorans DS-1	402881
65	Hyphomicrobium denitrificans ATCC 51888	582899
66	Methylobacterium nodulans ORS 2060	460265
67	Brucella abortus S19 89	430066

68	Agrobacterium tumefaciens F2	1050720
69	Polymorphum gilvum SL003B-26A1	991905
70	Candidatus Puniceispirillum marinum IMCC1322	488538
71	Micavibrio aeruginosavorus ARL-13	856793
72	Magnetospirillum magneticum AMB-1	342108
	Outgroup: Thermodesulfobium narugense DSM 14796	747365
	Outgroup: Streptococcus pneumoniae D39	373153
	Outgroup: Streptococcus gordonii str. Challis substr. CH1	467705
	Outgroup: Aerococcus urinae ACS-120-V-Col10a	866775
	Outgroup: Finegoldia magna ATCC 53516	525282
	Outgroup: Alicyclobacillus acidocaldarius subsp. acidocaldarius Tc-4-1	1048834
	Outgroup: Candidatus Desulforudis audaxviator MP104C	477974

d)	**Archaea** **Name**	**EggNogID**
1	Candidatus Nitrosoarchaeum koreensis MY1	1001994
2	Metallosphaera cuprina Ar-4	1006006
3	Halorhabdus tiamatea SARL4B	1033806
4	Thermococcus sp. 4557	1042877
5	Pyrobaculum sp. 1860	1104324
6	Methanosaeta harundinacea 6Ac	1110509
7	Pyrobaculum aerophilum str. IM2	178306
8	Pyrococcus furiosus DSM 3638	186497
9	Methanothermobacter thermautotrophicus str. Delta H	187420
10	Methanosarcina acetivorans C2A	188937
11	Methanopyrus kandleri AV19	190192
12	Methanosarcina mazei Go1	192952
13	Archaeoglobus fulgidus DSM 4304	224325
14	Nanoarchaeum equitans Kin4-M	228908
15	Methanocaldococcus jannaschii DSM 2661	243232
16	Thermococcus sp. AM4	246969
17	Methanococcoides burtonii DSM 6242	259564
18	Picrophilus torridus DSM 9790	263820
19	Methanococcus maripaludis S2	267377
20	Methanosarcina barkeri str. Fusaro	269797
21	Aeropyrum pernix K1	272557
22	Haloarcula marismortui ATCC 43049	272569
23	Pyrococcus abyssi GE5	272844
24	Sulfolobus solfataricus P2	273057
25	Sulfolobus tokodaii str. 7	273063
26	Thermoplasma acidophilum DSM 1728	273075
27	Thermoplasma volcanium GSS1	273116
28	uncultured marine group II euryarchaeote	274854
29	Methanocella paludicola SANAE	304371
30	Haloferax volcanii DS2	309800
31	Methanospirillum hungatei JF-1	323259
32	Sulfolobus acidocaldarius DSM 639	330779
33	Ferroplasma acidarmanus fer1	333146
34	Methanosphaera stadtmanae DSM 3091	339860
35	Pyrococcus sp. NA2	342949
36	Natronomonas pharaonis DSM 2160	348780

37	Methanosaeta thermophila PT	349307
38	Methanocella arvoryzae MRE50	351160
39	Haloquadratum walsbyi DSM 16790	362976
40	Methanoculleus marisnigri JR1	368407
41	Thermofilum pendens Hrk 5	368408
42	Candidatus Korarchaeum cryptofilum OPF8	374847
43	Pyrobaculum islandicum DSM 4184	384616
44	Thermococcus barophilus MP	391623
45	Caldivirga maquilingensis IC-167	397948
46	Metallosphaera sedula DSM 5348	399549
47	Staphylothermus marinus F1	399550
48	Methanococcus maripaludis C5	402880
49	Methanococcus vannielii SB	406327
50	Methanocorpusculum labreanum Z	410358
51	Pyrobaculum calidifontis JCM 11548	410359
52	Hyperthermus butylicus DSM 5456	415426
53	Halorubrum lacusprofundi ATCC 49239	416348
54	Methanococcus aeolicus Nankai-3	419665
55	Methanobrevibacter smithii ATCC 35061	420247
56	Methanococcus maripaludis C7	426368
57	Nitrosopumilus maritimus SCM1	436308
58	Aciduliprofundum boonei T469	439481
59	Pyrobaculum neutrophilum V24Sta	444157
60	Methanococcus maripaludis C6	444158
61	Ignicoccus hospitalis KIN4/I	453591
62	Methanococcus voltae A3	456320
63	Methanoregula boonei 6A8	456442
64	Halogeometricum borinquense DSM 11551	469382
65	Halomicrobium mukohataei DSM 12286	485914
66	Desulfurococcus kamchatkensis 1221n	490899
67	Halorhabdus utahensis DSM 12940	519442
68	Methanosphaerula palustris E1-9c	521011
69	Methanothermus fervidus DSM 2088	523846
70	Thermococcus onnurineus NA1	523850
71	Pyrococcus yayanosii CH1	529709
72	Haloterrigena turkmenica DSM 5511	543526
73	Methanohalophilus mahii DSM 5219	547558
74	Natrialba magadii ATCC 43099	547559
75	Vulcanisaeta distributa DSM 14429	572478
76	Archaeoglobus profundus DSM 5631	572546
77	Methanocaldococcus infernus ME	573063
78	Methanocaldococcus fervens AG86	573064
79	Methanocaldococcus vulcanius M7	579137
80	Ignisphaera aggregans DSM 17230	583356
81	Ferroglobus placidus DSM 10642	589924
82	Staphylothermus hellenicus DSM 12710	591019
83	Thermococcus gammatolerans EJ3	593117
84	Thermococcus sibiricus MM 739	604354
85	Thermosphaera aggregans DSM 11486	633148
86	Haloarcula hispanica ATCC 33960	634497
87	Methanobrevibacter ruminantium M1	634498
88	Halobacterium sp. NRC-1	64091

89	Methanocaldococcus sp. FS406-22	644281
90	Methanohalobium evestigatum Z-7303	644295
91	Methanothermococcus okinawensis IH1	647113
92	Acidilobus saccharovorans 345-15	666510
93	Methanosalsum zhilinae DSM 4017	679901
94	Methanoplanus petrolearius DSM 11571	679926
95	Thermococcus kodakarensis KOD1	69014
96	Archaeoglobus veneficus SNP6	693661
97	Pyrolobus fumarii 1A	694429
98	Pyrobaculum oguniense TE7	698757
99	Pyrococcus horikoshii OT3	70601
100	halophilic archaeon DL31	756883
101	Desulfurococcus mucosus DSM 2162	765177
102	Thermoproteus tenax Kra 1	768679
103	Halalkalicoccus jeotgali B3	795797
104	Haladaptatus paucihalophilus DX253	797209
105	Halopiger xanaduensis SH-6	797210
106	Methanothermobacter marburgensis str. Marburg	79929
107	Methanobacterium paludis	868131
108	Methanobacterium lacus	868132
109	Methanotorris igneus Kol 5	880724
110	Candidatus Nitrosoarchaeum limnia SFB1	886738
111	Sulfolobus islandicus REY15A	930945
112	Acidianus hospitalis W1	933801
113	Vulcanisaeta moutnovskia 768-28	985053
114	Methanosaeta concilii GP6	990316
115	Thermoproteus uzoniensis 768-20	999630