bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Phylogenomic Testing of Root Hypotheses Without a Species Tree

Fernando D. K. Tria[1], Giddy Landan[1*], Tal Dagan

Genomic Microbiology Group, Institute of General Microbiology, Kiel University, Kiel, Germany

[1] Equally contributed.

[*] Corresponding author.

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ABSTRACT

The determination of the last common ancestor (LCA) of a group of species plays a vital role in evolutionary theory. Traditionally, an LCA is inferred by the rooting of a fully resolved species tree. From a theoretical perspective, however, inference of the LCA amounts to the reconstruction of just one branch - the root branch - of the true unrooted species tree, and should therefore be a much easier task than the full resolution of the species tree. Discarding the reliance on a hypothesised species tree and its rooting leads us to re-evaluate what phylogenetic signal is directly relevant to LCA inference, and to recast the task as that of sampling the total evidence from all gene families at the genomic scope. Here we reformulate LCA and root inference in the framework of statistical hypothesis testing and outline an analytical procedure to formally test competing *a-priori* LCA hypotheses and to infer confidence sets for the earliest speciation events in the history of a group of species. Applying our methods to three demonstrative datasets we show that our inference of the metazoan LCA as well as the cyanobacterial LCA is well in agreement with the common knowledge. Inference of the proteobacteria LCA shows that it is most closely related to modern Epsilonproteobacteria, hence it was most likely characterized by a chemolithoautotrophic and anaerobic life-style. Our inference is based on data comprising between 43% (opisthokonta) and 86% (proteobacteria) of all gene families. Approaching LCA inference within a statistical framework thus renders the phylogenomic inference powerful and robust.

Keywords: Phylogenetics, Species tree, Last Common Ancestor (LCA), Cyanobacteria, Proteobacteria.

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Inference of the last common ancestor (LCA) for a group of taxa is central for the study of evolution of genes, genomes and organisms. Discoveries and insights from the reconstruction of LCAs span a wide range of taxonomic groups and time scales. For example, studies of the last universal common ancestor (LUCA) of all living organisms inferred that it was an anaerobic organism whose energy metabolism was characterized by CO_2-fixing, H_2-dependent with a Wood–Ljungdahl pathway, N_2-fixing and thermophilic (Weiss et al. 2016). Nonetheless, due to the inherent difficulty of ancient LCA inference, those are frequently the centre of evolutionary controversies, such as for instance, the debate concerning the two versus three domains of life (Williams et al. 2013), the LCA of vertebrates (Okamoto et al. 2017), or the LCA of hominids (Lovejoy et al. 2009).

The identity of the LCA is traditionally inferred from a species tree that is reconstructed unrooted and is then rooted at the final step. Thus, the LCA inference is dependent on the accuracy of the species tree topology. One approach for the reconstruction of a species tree, is to use a single gene as a proxy for the species tree topology, e.g., 16S ribosomal RNA subunit for prokaryotes (Fox et al. 1980) or the Cytochrome C for eukaryotes (Fitch and Margoliash 1967). This approach is, however, limited in its utility due to possible differences between the gene evolutionary history and the species phylogeny. Examples are incongruence due to lateral gene transfer in bacteria evolution (Dickerson 1980), endosymbiotic gene transfer in eukaryotic evolution (Martin et al. 2002), or hybridization events in plants (Velasco et al. 2010). Phylogenomics offer an alternative to the single-gene approach as this approach aims to utilize the whole genome rather then a single gene for the phylogenetic reconstruction (Eisen and Fraser 2003). In the most basic approach, the species tree is reconstructed from the genes that are shared among all the species under study, termed here as complete gene families. These genes can be used for the

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reconstruction of a species tree using several approaches including tree reconstruction from concatenated alignments (e.g., Ciccarelli et al. 2006; Parks et al. 2018) as well as consensus trees (e.g., Dagan et al. 2013). However, these approaches are often restricted in their data sample as they exclude partial gene families that are not present in all members of the species set, e.g., due to differential loss. Furthermore, methods based on complete gene families are restricted to single-copy gene families; hence they exclude multi-copy gene families that are present in multiple copies in one or more species, e.g., as a result of gene duplications or gene acquisition. This is because the evolution of multi-copy gene families differs from that of the species tree. Thus the drawback in alignment concatenation or consensus tree approaches is that the inference becomes limited to gene sets that do not represent the entirety of genomes. This issue tends to become more acute the more diverse the species set is. In extreme cases no single-copy, complete gene family exists (Medini et al. 2005). Super-trees approaches offer an alternative as they enable to include also partial gene families (Pisani et al. 2007; Whidden et al. 2014; Williams et al. 2017); however, those approaches also exclude multi-copy gene families (partial and complete). Thus, while the major aim of phylogenomics approaches is to improve the accuracy of phylogenetic inference by increasing the sample size, all current methodologies suffer from several inference problems, with some elements in the formulation that are common to all of them. The first is the limited sample sizes due to the number of single copy genes. That is the tree of 1% (Dagan and Martin 2006, and see our Table 1), the existence of reduced genomes, and the spanning of taxon groups that in the extreme case have no common single copy genes. In both approaches there is no room for the inclusion of families including paralogous genes, and furthermore, reduction of families including paralogs into orthologs-only subsets, e.g., using tree reconciliation (Szöllosi et al. 2015), requires an assumed species tree topology. Finally, since the aforementioned approaches

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

yield unrooted species trees, the inference of the root is performed as the last step in the analysis, hence the sample size for the LCA inference is essentially a single tree.

The inference of the LCA from a single species tree can be robust and accurate only if the underlying species tree is reliable. Unfortunately, this is rarely the case, as can be frequently seen in the plurality of gene tree topologies and their disagreement with species trees (e.g. Doolittle and Bapteste 2007; Linz et al. 2007, and see our Fig. 5a). We propose that for the identification of an LCA, one does not need to reconstruct a fully resolved species tree. Instead, the LCA can be defined as the first speciation event for the group of species. In this formulation, the topological resolution of the entire species tree is immaterial and the only phylogenetic conclusion needed is the partitioning of the species into two groups. Here we present a novel approach for the inference of the LCA without reconstructing a species tree. Our approach considers the total evidence from unrooted gene trees for all protein families from a set of taxa, including partial families as well as those with paralogous gene copies.

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MATERIALS & METHODS

We present the rooting approach with the help of illustrative rooting problems for three species sets: opisthokonta, cyanobacteria, and proteobacteria (Table 1). The root position is well established for the opisthokonta and cyanobacteria species sets, and they serve here as positive controls. The root of the proteobacteria species set is still debated, and it serves to demonstrate the power of the proposed procedure. The opisthokonta dataset comprises 14 metazoa and 17 fungi species, and the known root is a partition separating fungi from metazoa species (Stechmann and Cavalier-Smith 2002; Katz et al. 2012). The cyanobacteria dataset spans five morphological sections, and the root partition separates 31 marine unicellular species from the others (unicellular and multicellular species) (Tria et al. 2017). The proteobacteria dataset includes species from five taxonomical classes in that phylum (Ciccarelli et al. 2006; Pisani et al. 2007; Lang et al. 2013, but see Waite et al. 2017). The Proteobacteria dataset poses a harder root inference challenge than cyanobacteria and ophistokonta as previous results suggest the existence of a root neighborhood of three competing branches (Tria et al. 2017). Protein families for the opisthokonta 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) (Table 1).

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. Illustrative datasets, their consensus MAD rooting, and classification of gene families. For the complete list of species, see Supplementary Table 1.

	Opisthokonta	**Cyanobacteria**	**Proteobacteria**
Number of species	31	130	172
Consensus MAD Root in CSC gene trees	77.70% (132 trees) {Fungi , Metazoa}	72.60% (83.5 trees) {SynPro clade, other cyanobacteria}	33.17% (16.583 trees) {ε-proteobacteria, other proteobacteria}
Number of gene families	13036	17918	9686
CSC: Complete single-copy gene families, present as single-copy in all members of a species set.	170 (1.30%)	115 (0.64%)	50 (0.52%)
CMC: Complete multi-copy gene families, present in all species, but having multiple copies in at least one species.	612 (4.69%)	57 (0.32%)	70 (0.72%)
PSC: Partial single-copy gene families, absent from some species and present as single-copy in the others.	7773 (59.63%)	13321 (74.34%)	5586 (57.67%)
PMC: Partial multi-copy gene families, absent from some species and having multiple copies in at least one other species.	4481 (34.37%)	4425 (24.70%)	3980 (41.09%)

Protein families were filtered based on the number of species, gene copy number, number of OTUs, and sequence length, as follows. Protein families present in less than four species were discarded. Suspected outlier sequences were detected based on their length relative to the median length: sequences were removed if shorter than half or longer

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

than twice the median. Species with more than ten copies of a gene were removed from the corresponding gene family. Multi-copy gene families were discarded if the number of species was smaller than half the total number of OTUs (Table 1).

Protein sequences of the resulting protein families were aligned using MAFFT version 7.027b with L-INS-i alignment strategy (Katoh and Standley 2013). Phylogenetic trees were reconstructed using iqtree version 1.6.6 with the model selection parameters '-mset LG -madd LG4X' (Nguyen et al. 2015). The phylogenetic network (Fig. 5) was reconstructed using SplitsTree4 version 4.14.6 (Huson and Bryant 2006). Branch AD values and roots for the consensus analysis were inferred using mad.py version 2.21 (Tria et al. 2017).

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Terms and Definitions

LCA inference deals with abstractions of similar but distinct types of trees: the hypothetical true *species tree* and *gene trees*. It is therefore helpful clearly to delineate the terms we use and the sense in which we use them.

OTUs (Operational Taxonomic Units) – The leafs of a gene tree. A species may be represented by multiple OTUs in a gene tree.

Split – The OTU bipartition induced by a branch in a phylogenetic tree.

Species split – A split where all OTUs of any one species are present on the same side. In a species tree, all splits are species splits.

Root – The deepest internal node in a rooted phylogenetic tree (either gene or species tree), representing the last common ancestor (LCA) of all the OTUs.

Root branch – The branch in an unrooted phylogenetic tree that harbours the root node.

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

LCA – Last common ancestor. Here we restrict this term to the context of species. The LCA is the root node of a species tree.

LCA partition = Species root partition – We reserve this term for species *sets*, without reference to a particular species tree (hence the use of partition, and not split). It represents the immediate diversification of the LCA into two lineages. In a hypothetical *species tree*, it is identical to the root (species) split.

LCA confidence set = Root neighbourhood – Multiple, equally likely species root partitions for a species set.

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RESULTS

Our LCA inference approach differs from existing ones in several aspects: 1) No species tree is reconstructed or assumed. 2) Phylogenetic information is extracted from gene trees reconstructed from partial and multi-copy gene families in addition to CSC gene families. 3) The analysis uses unrooted gene trees and no rooting operations are performed, for either gene trees or species trees. 4) Any LCA hypothesis can be tested, including species partitions that do not occur in any of the gene or species trees. Before describing our approach, we first demonstrate the limitations of a simpler phylogenomic rooting procedure that uses CSC gene families and infers the root by a consensus derived from the rooted trees of the CSC genes. We then show how to extract additional information from unrooted gene trees. The incorporation of additional information not considered by a simple consensus of rooted trees leads to a statistical test to decide between two competing root hypotheses. Next we show how information from partial and multi-copy gene families can be used within the same statistical framework, greatly increasing the sample size and inference power. We then extend the pairwise formulation and consider multiple competing root partitions. Finally, we modify the pairwise test to a one-to-many test, and present a sequential elimination process that infers a minimal root neighbourhood, i.e., a confidence set of LCA partitions.

Phylogenomic consensus rooting

The consensus approach infers the root partition of a species set from a sample of rooted CSC gene trees. Root splits are collected from all trees and the most frequent root split is the inferred species root partition for the species set. In species sets with a strong root signal, this majority-rule approach is sufficient to determine a clear root partition for the species set. This circumstance is observed in the opisthokonta and cyanobacteria illustrative datasets. Using MAD (Tria et al. 2017; Bryant and Charleston 2018) to root the

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

individual gene trees, the consensus species root partition was inferred as the root split in more than 70% of the CSC gene trees, in both datasets (see Table 1). In the proteobacteria, in contrast, the most frequent root branch was inferred in 33% of the CSC gene trees, while two competing root braches are observed in almost 15% of the gene trees. The performance of the consensus approach is thus hindered by three factors. First, majority-rule voting considers just one split from each gene tree, ignoring a large measure of the phylogenetic signal present in the gene trees. In addition, the quality of the root inference varies among the gene trees and is quantifiable, but this information is not utilized by the consensus approach. Lastly, simple voting cannot be satisfactorily tested for statistical significance.

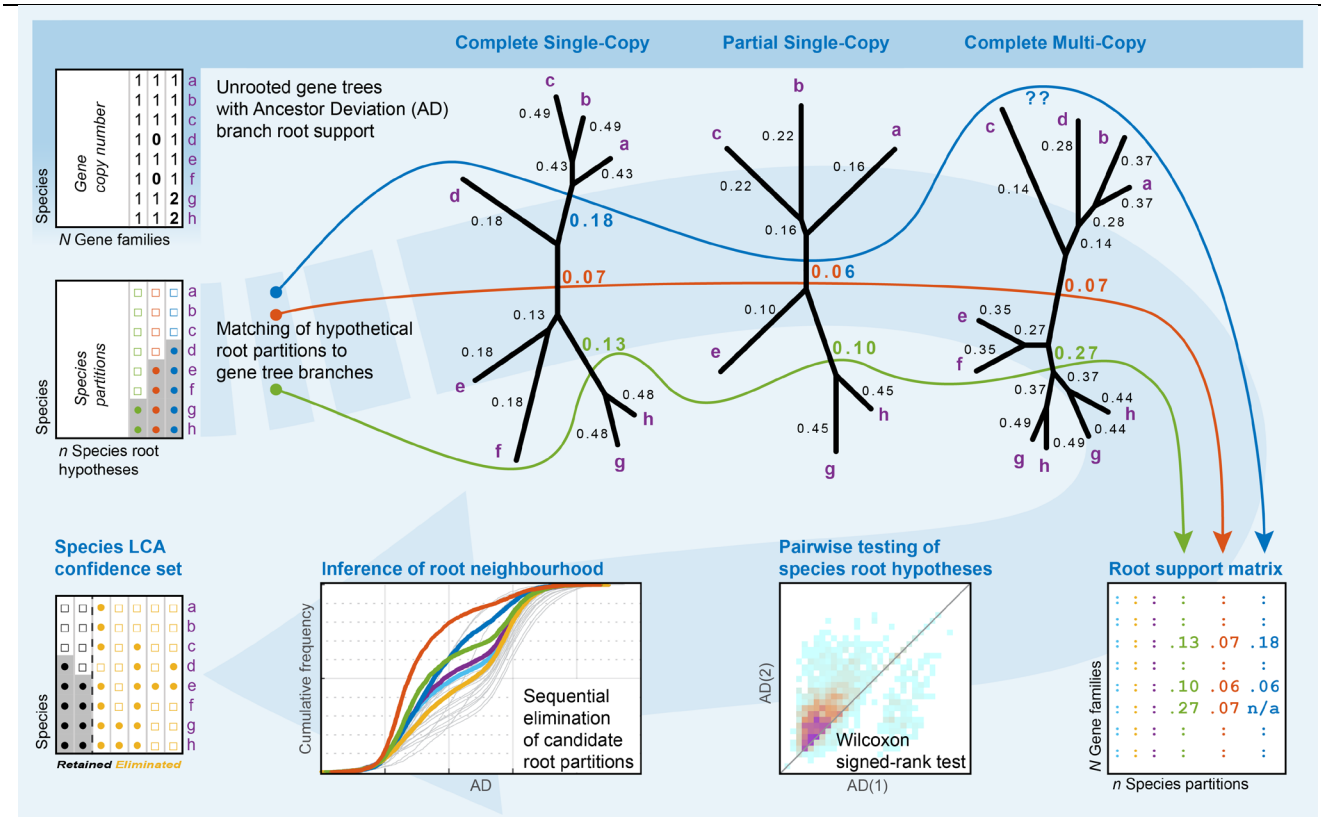


Figure 1. Outline of the analytical procedure. Stages are depicted clockwise from top-left.

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The root support test for two alternative root partitions

The first step in our approach is a formulation of a test to select between two competing species root partitions (see Fig. 1 for a road-map of the procedure). The test considers the Ancestor Deviation (AD) statistic for the competing root hypotheses in individual gene trees. The AD measure quantifies the amount of lineage rate heterogeneity that is induced by postulating a branch as harbouring the root of the tree. We have previously shown that the AD measure provides robust evidence for the inference of the root of a single gene tree (Tria et al. 2017). In the current study, we do not infer a single root for each gene tree, but use the AD measure to assess the relative strength of alternative rootings of the same tree. Collecting AD values from a set of gene trees, we obtain a paired sample of support values. In Fig. 2a we present the joint distribution of AD values for the two most likely root partitions in the eukaryotic dataset. A null hypothesis of equal support can be tested by the Wilcoxon signed-rank test, and rejection of the null hypothesis indicates that the root partition with smaller AD values is significantly better supported than the competitor.

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

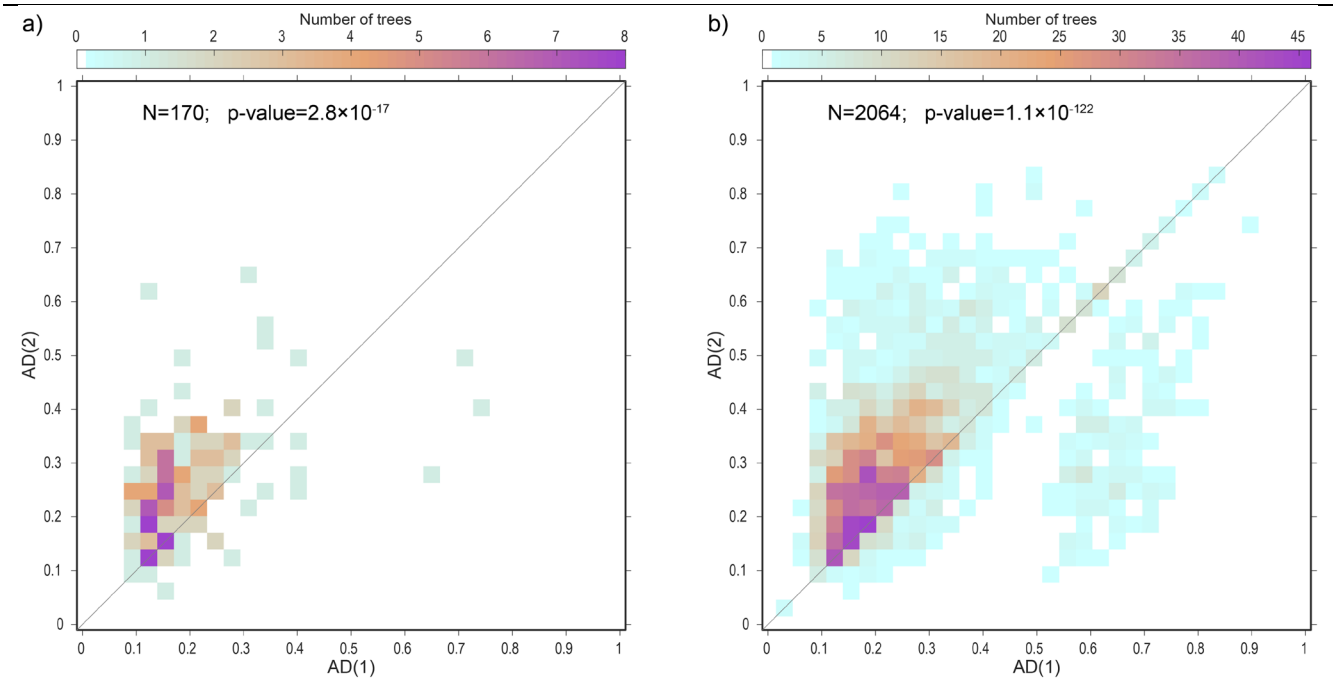


Figure 2. Pairwise testing of competing root hypotheses in the opisthokonta dataset. **a)** CSC gene families; **b)** All gene families. Colormaps are the joint distribution of paired AD values. Candidates 1 and 2 are the two most frequent root partitions among the CSC gene trees (Supplementary Table 1a). Smaller ADs indicate better support, whereby candidate 1 out-compete candidate 2 above the diagonal and candidate 2 wins below the diagonal. P-values are for the two-sided Wilcoxon signed-rank test. Note the gain in power concomitant to larger sample size.

As in all statistical inferences, the power of the test ultimately depends on the sample size. Considering only CSC gene families often limits rooting analyses to a small minority of the available sequence data (e.g., Table 1). Paired AD support values, however, can be extracted also from partial and paralogous gene families, resulting in much larger sample size and statistical power (Fig. 2b).

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Rooting support from partial and multi-copy gene trees

In CSC gene trees the correspondence between branches and hypothesized root partitions is direct and one-to-one (Fig 3a). To deal with non-CSC gene trees we must decouple the notion of a 'tree branch' or ‘split‘ from that of a 'root partition'. In trees of partial gene families, a single branch may correspond to several species root partitions. In multiple-copy gene families, some tree branches do not correspond to any possible species root partition.

In order to find the branches in a *partial* gene tree that correspond to the root partitions, we *reduce* root partitions from species to OTUs by removing the species that are missing in the gene tree (Semple and Steel 2000). The root partitions are then assigned AD support by matching their reduced OTU version to the OTU splits of the gene tree (Fig 3b).

In *multi-copy* gene trees one or more species are represented multiple times as an OTU (Swenson and El-Mabrouk 2012). Each branch of a multi-copy gene tree splits the OTUs into two groups, and the two groups may be mutually exclusive or overlapping in terms of species. Species splits (i.e., mutually exclusive) can be mapped to specific root partitions. Overlapping splits, on the other hand, cannot correspond to any root partition (Figure 3c). Mapping of tree splits from partial multi-copy gene trees entail both operations: identification of species splits and reduction of root partitions.

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

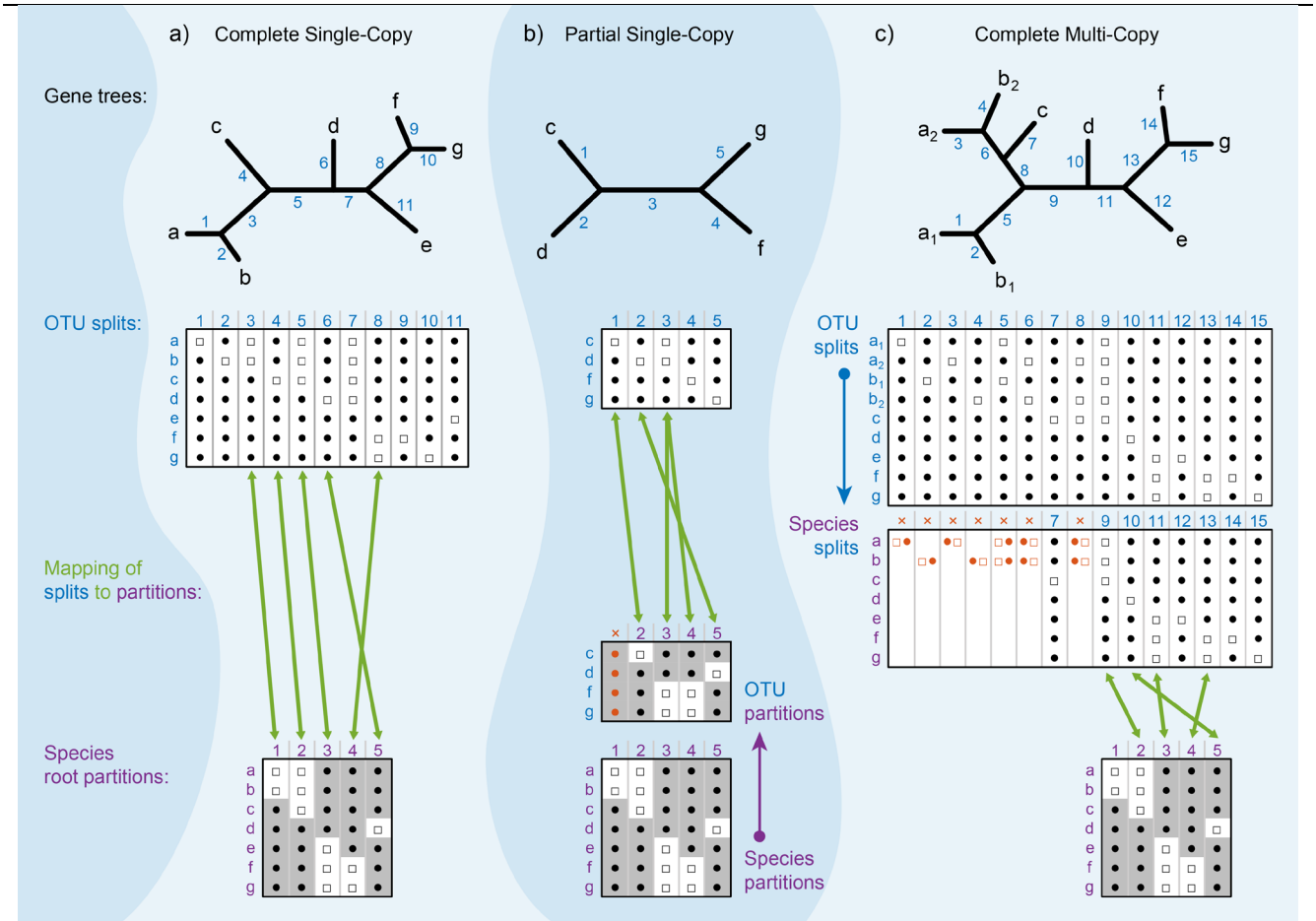


Figure 3. Correspondence of OTU splits and root partitions. a) In CSC gene trees; **b)** In PSC gene trees; c) In CMC gene trees. PMC gene trees entail both the **a)** and **b)** operations.

Candidate root partitions, or their reduced versions, may be absent from some gene trees, and will be missing support values from these trees. We distinguish between two such cases: informative and uninformative missing values. A gene family is uninformative relative to a species root partition when its species composition includes species from only one side of the species partition. In such cases, the candidate root partition cannot be observable in any reconstructed gene tree. We label the gene trees of such families as uninformative relative to the candidate root partition, and exclude them from tests involving

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

that partition. In contrast, when a gene family includes species from both sides of a candidate species root partition but the gene tree lacks a corresponding branch, we label the gene tree as informative relative to the partition. This constitutes evidence against the candidate partition, and should not be ignored in the ensuing tests. In such cases we replace the missing support values by a pseudo-count consisting of the maximal (i.e., worst) AD value in the gene tree. This assignment of a default worst-case support value also serves to enable the pairwise testing of incompatible root partitions, where no gene tree can include both partitions (Semple and Steel 2000).

Complete gene families are always informative relative to any candidate root partitions. Partial gene families, however, may be uninformative for some root candidates. When testing two candidate root hypotheses against each other, the exclusion of uninformative partial gene trees thus leads to a reduction of sample size from the full complement of gene families. Furthermore, one branch of a partial gene tree may be identical to the reduced versions of two or more species root partitions, whereby the tree is informative relative to the several candidates yet their support value is tied.

Root inference and root neighborhoods

The pairwise test is useful when the two competing root hypotheses are given *a-priori*, as often happens in specific evolutionary controversies. More often, however, one wishes to infer the species LCA, or root partition, with no prior hypotheses. In principle, the pairwise test may be carried out over all pairs of possible root partitions, while controlling for multiple testing. Such an exhaustive approach is practically limited to very small rooting problems, as the number of possible partitions grows exponentially with the number of species. A possible simplification is to restrict the analysis to test only pairs of root partitions from a pool of likely candidates. We propose that a reasonable pool of candidate root partitions

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

can be constructed by collecting the set of root splits that are inferred as the root in any of the CSC gene trees.

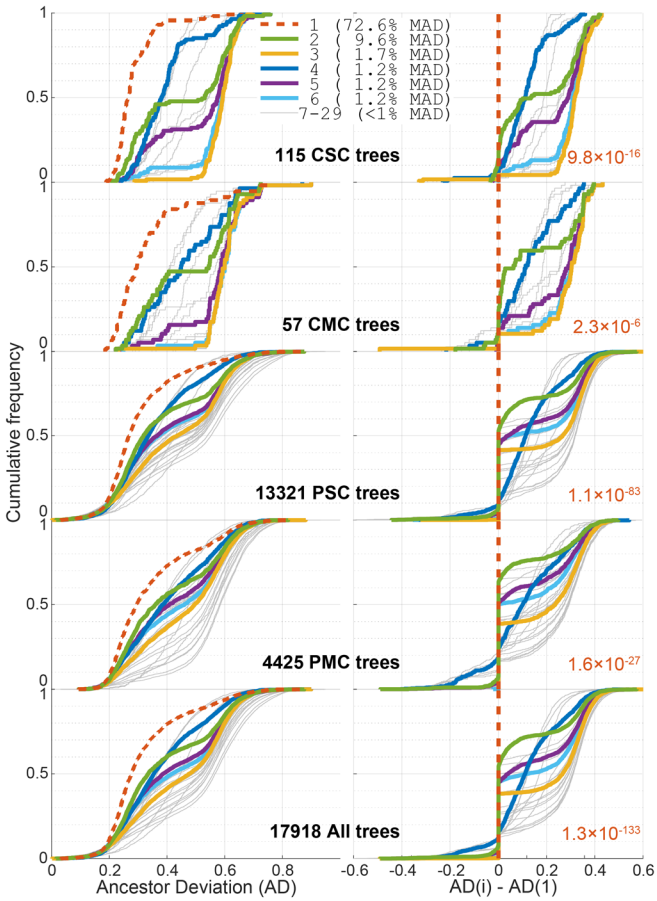


Figure 4. Cumulative distribution plots of AD in the cyanobacteria dataset. Vertically are stacked the subsets by gene family type. On the left are unpaired AD values for the 29 candidate root partitions. On the right are the paired differences to candidate 1, whereby positive differences indicate better support for candidate 1 and negative values better support for candidate 2. (See Supplementary Table 1b for candidate partition definitions). In red are p-values of the least significant among the contrasts to candidate 1, FDR adjusted for all 406 pairwise comparisons (Supplementary Table 2b). Note the overall similarity between the different subsets, indicating a common and robust root signal.

When one species root partition is significantly better supported than any of the other candidates, the root is fully determined. Such is the result for the opisthokonta and cyanonacteria datasets, for which the known root partition is the best candidate among all pairwise comparisons (Figure 4 and Supplementary Table 2). In more difficult situations the interpretation of all pairwise *p*-values is not straightforward due to the absence of a

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

unanimous best candidate root partition. This situation is exemplified with the CSC subset of the proteobacteria dataset where no candidate has better support than all the alternative candidates (Supplementary Table 2c). The absence of a clear best candidate suggests the existence of a root neighbourhood in the species set. Thus, a rigorous procedure for the inference of a confidence set for LCA is required.

One-to-many root support test

To assess the support for root partitions in the full context of all other candidate root partitions, we modify the pairwise test to a test contrasting one root partition to a set of many alternatives. The *One-to-Many test* consists of comparing the distribution of root support values for one focal partition to the extreme support values among all the other candidates, and is inherently asymmetric. A 'Better than Best' version takes the minimal (i.e., best) value among the AD values of the alternatives, while the 'Worse than Worst' version considers the maximal (i.e., worst) among the alternatives' ADs. As expected, the 'better than best' variant is always less powerful than any of the pairwise tests, and will not be considered further. The 'worse than worst' variant, on the other hand, can be used to trim down a set of candidates while being more conservative than the pairwise tests. In the one-to-many test, each gene tree provides one AD value for the focal partition and one AD value for the worst among the alternative root partitions. Note that the worst alternative root partition may vary across gene trees. We test for differences in the magnitude of paired AD values using the one-sided Wilcoxon-signed rank test, with the null hypothesis that the focal ADs are equal or smaller than the maximal ADs for the complementary set, and the alternative hypothesis that the focal ADs are *larger still* than the maximum. A rejection of the null hypothesis is interpreted to mean that the focal root partition is significantly worse supported than the complementary set of candidates taken as a whole.

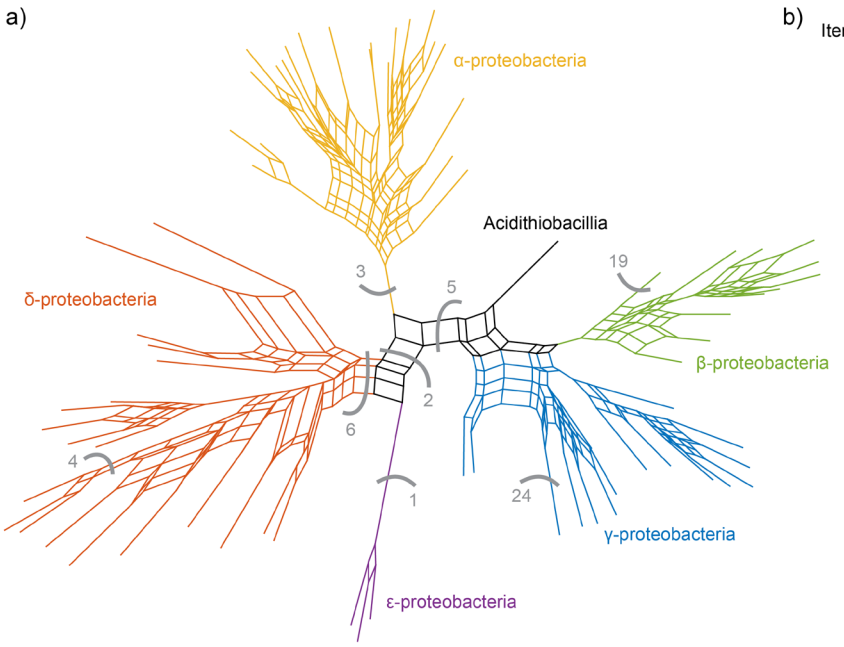
bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Inference of a minimal neighbourhood

To infer a root neighbourhood, i.e., a confidence set of LCA hypotheses, we start with a reasonably constructed large set of *n* candidate partitions, and reduce it by a stepwise elimination procedure. At each step, we employ the one-to-many test to contrast each of the remaining candidates to its complementary set. We control for multiple testing using FDR (Benjamini and Hochberg 1995), and if at least one test is significant at the specified FDR level, the focal partition with the smallest p-value (i.e., largest z-statistic) is removed from the set of candidates. The iterative process is stopped when none of the retained candidates is significantly worse supported than the worst support for the other members of the set, or when the set is reduced to a single root partition. To be conservative, we use a cumulative FDR procedure where at the first step we control for *n* tests, in the next round for $2n-1$ tests, and, when not stooped earlier, for $n*(n-1)/2-1$ at the last iteration.

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a)



b)

Iteration	Rejected partition	MAD root in CSC	Number of trees	FDR adjusted p-value
1	23	0.5%	7440	5.0×10^{-228}
2	16	2.0%	4022	2.8×10^{-204}
3	15	2.0%	7395	4.5×10^{-298}
4	9	2.0%	6916	2.2×10^{-240}
5	11	2.0%	6969	2.9×10^{-233}
6	12	2.0%	6516	1.6×10^{-214}
7	22	0.5%	6043	2.5×10^{-228}
8	**5**	5.1%	6762	5.9×10^{-222}
9	21	0.5%	5933	6.0×10^{-220}
10	10	2.0%	6457	5.0×10^{-211}
11	13	2.0%	5759	2.6×10^{-207}
12	18	1.0%	6818	1.0×10^{-198}
13	25	0.4%	5656	6.1×10^{-196}
14	8	2.0%	5481	2.7×10^{-194}
15	14	2.0%	5613	9.5×10^{-180}
16	17	1.0%	5232	1.5×10^{-182}
17	20	0.7%	5464	5.5×10^{-166}
18	**6**	3.3%	5432	5.4×10^{-144}
19	**3**	12.1%	4074	4.4×10^{-92}
20	**2**	14.7%	3489	8.2×10^{-137}
21	**4**	6.0%	1764	3.5×10^{-93}
22	7	2.0%	2363	3.2×10^{-60}
23	**24**	0.4%	938	4.8×10^{-24}
24	**19**	0.7%	664	1.5×10^{-4}
Retained	**1**	33.2%		

Figure 5. LCA inference by sequential elimination in the proteobacteria dataset.

a) Phylogenetic split network of the 50 CSC gene trees; **b)** Trace of the sequential elimination process (See Supplementary Table 1c for candidate partition definitions). Selected partitions are indicated by grey arcs in **a)** and bold numbers in **b)**.

We demonstrate the sequential elimination procedure for the proteobacteria dataset in figure 5. The splits network reconstructed for the proteobacteria dataset exemplifies the plurality of incongruent splits in the CSC gene trees and hence the dangers in assuming a single species tree. In this dataset the initial candidate set consisted of the 25 different root partitions found in the 50 CSC gene trees, and the elimination process terminated with a neighbourhood of size 1, a species root partition separating the Epsilonproteobacteria species from the other proteobacteria classes. This LCA is indeed the most frequent one among the CSC gene trees, but with a low frequency of only one in three gene trees. It is noteworthy that the order of elimination does not generally follow the frequency of partitions in the CSC set. For example, the last alternative to be rejected (number 19) was inferred as

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a root branch in just one tree where it is tied with two other branches, whereas the second and third most frequent CSC roots are rejected already at iterations 19-20.

The elimination order is determined by the p-value of the one-to-many test, which in turn reflects both the effect size of worse support and the power of the test, where the latter is a function of sample size. Hence, candidate partitions for which a smaller number of gene trees are informative are more difficult to reject. In particular, the testing of an LCA hypothesis of a single basal species partitioned from the other species is limited to those gene families that include the basal species. The last two partitions rejected in Figure 5 are indeed single species partitions, and the number of gene trees that are informative relative both to them and to the remaining candidates drops drastically in comparison to earlier iterations. Yet, even at the last iteration the number of gene trees that bear upon the conclusion is an order of magnitude larger than the number of CSC gene families.

The full complement of the proteobacteria dataset consists of 9,686 gene families. The final conclusion - determination of a single LCA partition - is arrived at by extracting ancestor-descendant information from 86% of the gene families. The gene families that do not provide any evidence consist of 1113 PSC gene families, mainly very small ones (e.g., due to recent gene origin), and 214 PMC families, mostly small families and some with abundant paralogs (e.g., due to gene duplication prior to the LCA). A fundamental element in our approach is the prior definition of a pool of candidate root partitions. We advocate deriving the initial set from roots inferred for CSC gene trees. A yet larger but manageable initial set may constructed of splits frequently observed in the CSC gene trees. Importantly, the initial set need not be limited to observed partitions, but can be augmented by *a-priori* hypotheses informed by current phylogenetic and taxonomical percepts.

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DISCUSSION

The inferred species root partition for the proteobacteria dataset indicates that the proteobacteria LCA was more closely related to modern Epsilonproteobacteria in comparison to the other classes; characteristics of present-day species in that group can therefore be used to hypothesize about the biology of the proteobacterial LCA. 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)). The class includes in addition several pathogenic organisms (e.g., *Campylobacter*, *Helicobacter*) that are associated with Human or other mammals; yet its LCA was likely a thermophilic species inhabiting deep-sea hydrothermal vents (Zhang and Sievert 2014). Epsilonproteobacteria residing in deep-sea habitats 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). Members of the Epsilonproteobacteria have been shown to thrive in deep sea hydrothermal vents, most of those are characterized as chemolithoautotrophs (Takai et al. 2005). Taken together, these observations, and the epsilonproteobacteria root, suggest that: 1) the proteobacteria LCA was an 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).

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

From a purely theoretical perspective, the inference of the LCA for a group of species amounts to the reconstruction of just one branch - the root branch - of the true unrooted species tree, and should therefore be a much easier task than the full resolution of the rooted species tree. Traditional approaches, however, posed the LCA problem in terms of rooting of a resolved species tree, a formulation with two major drawbacks. First, it requires the solution of a much harder problem as a prerequisite for addressing the easier task. Secondly, the input information passes through a bottleneck of a single inferred species tree, so that the actual inference of the LCA is based on a sample of size one.

Avoiding the reliance on a species tree prompt us to re-evaluate what phylogenetic signal is directly relevant to LCA inference, and to recast the task as that of sampling the total evidence from all gene families at the genomic scope. Moreover, dispensing with a single rooting operation of a single species tree facilitates the reformulation of LCA and root inference in the framework of statistical hypothesis testing. The analytical procedure we outline allows formally to test competing *a-priori* LCA hypotheses and to infer confidence sets for the earliest speciation events in the history of a group of species.

Our analyses of the demonstrative datasets show that different species sets present varying levels of LCA signal: the opitsthokonta and cyanobacteria datasets show a strong root signal, the proteobacteria dataset has a moderate LCA signal. Datasets with weak signal are better described in terms of a confidence sets for root partitions, reflecting the inherent uncertainties and avoiding the pitfalls in forcing a single-hypothesis result.

The LCA inferences presented here utilized 43-86% of the total number of gene families for root partition inferences. This is in stark contrast to the 0.5-1.3% of the gene families that are CSC and can be utilized by traditional approaches. In the most extreme cases, the inclusion of non-CSC gene trees paves the way for root inferences in datasets

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with no complete gene families. The number of genes families considered in our tests corresponds to the number of genes encoded in modern genomes, supplying 'total evidence' for LCA inferences.

DATA AVAILABILITY

Waiting for a dryad URL.

REFERENCES

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

Benjamini Y., Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B Methodological. 57:289–300.

Bryant D., Charleston M. 2018. MAD roots for large trees, https://arxiv.org/abs/1811.03174v1

Campbell B.J., Engel A.S., Porter M.L., Takai K. 2006. The versatile epsilon-proteobacteria: key players in sulphidic habitats. Nat. Rev. Mol. Cell Biol. 4:458–468.

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

Dagan T., Martin W. 2006. The tree of one percent. Genome Biol. 7:118.

Dagan T., Roettger M., Stucken K., Landan G., Koch R., Major P., Gould S.B., Goremykin V.V., Rippka R., Tandeau de Marsac N., Gugger M., Lockhart P.J., Allen J.F., Brune I.,

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Maus I., Pühler A., Martin W.F. 2013. Genomes of Stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. Genome Biol. Evol. 5:31–44.

Dickerson R.E. 1980. Evolution and gene transfer in purple photosynthetic bacteria. Nature. 283:210–212.

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

Eisen J.A., Fraser C.M. 2003. Phylogenomics: intersection of evolution and genomics. Science. 300:1706–1707.

Fitch W.M., Margoliash E. 1967. Construction of phylogenetic trees. Science. 155:279–284.

Fox G.E., Stackebrandt E., Hespell R.B., Gibson J., Maniloff J., Dyer T.A., Wolfe R.S., Balch W.E., Tanner R.S., Magrum L.J., Zablen L.B., Blakemore R., Gupta R., Bonen L., Lewis B.J., Stahl D.A., Luehrsen K.R., Chen K.N., Woese C.R. 1980. The phylogeny of prokaryotes. Science. 209:457–463.

Huerta-Cepas J., Szklarczyk D., Forslund K., Cook H., Heller D., Walter M.C., Rattei T., Mende D.R., Sunagawa S., Kuhn M., Jensen L.J., Mering von C., Bork P. 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. 44:D286–93.

Huson D.H., Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254–267.

Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7:

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

435 improvements in performance and usability. Mol. Biol. Evol. 30:772–780.

436 Katz L.A., Grant J.R., Parfrey L.W., Burleigh J.G. 2012. Turning the crown upside down:
437 gene tree parsimony roots the eukaryotic tree of life. Systematic Biol. 61:653–660.

438 Lang J.M., Darling A.E., Eisen J.A. 2013. Phylogeny of Bacterial and Archaeal Genomes
439 Using Conserved Genes: Supertrees and Supermatrices. PLoS ONE. 8:e62510–15.

440 Linz S., Radtke A., Haeseler von A. 2007. A likelihood framework to measure horizontal
441 gene transfer. Mol. Biol. Evol. 24:1312–1319.

442 Lovejoy C.O., Suwa G., Simpson S.W., Matternes J.H., White T.D. 2009. The great divides:
443 Ardipithecus ramidus reveals the postcrania of our last common ancestors with African
444 apes. Science. 326:100–106.

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

447 Martin W., Rujan T., Richly E., Hansen A., Cornelsen S., Lins T., Leister D., Stoebe B.,
448 Hasegawa M., Penny D. 2002. Evolutionary analysis of Arabidopsis, cyanobacterial,
449 and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial
450 genes in the nucleus. Proc. Natl. Acad. Sci. U.S.A. 99:12246–12251.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

O'Leary N.A., Wright M.W., Brister J.R., Ciufo S., Haddad D., McVeigh R., Rajput B., Robbertse B., Smith-White B., Ako-Adjei D., Astashyn A., Badretdin A., Bao Y., Blinkova O., Brover V., Chetvernin V., Choi J., Cox E., Ermolaeva O., Farrell C.M., Goldfarb T., Gupta T., Haft D., Hatcher E., Hlavina W., Joardar V.S., Kodali V.K., Li W., Maglott D., Masterson P., McGarvey K.M., Murphy M.R., O'Neill K., Pujar S., Rangwala S.H., Rausch D., Riddick L.D., Schoch C., Shkeda A., Storz S.S., Sun H., Thibaud-Nissen F., Tolstoy I., Tully R.E., Vatsan A.R., Wallin C., Webb D., Wu W., Landrum M.J., Kimchi A., Tatusova T., DiCuccio M., Kitts P., Murphy T.D., Pruitt K.D. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. 44:D733–45.

Okamoto E., Kusakabe R., Kuraku S., Hyodo S., Robert-Moreno A., Onimaru K., Sharpe J., Kuratani S., Tanaka M. 2017. Migratory appendicular muscles precursor cells in the common ancestor to all vertebrates. Nat. Ecol. Evol. 1:1731–1736.

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

Pisani D., Cotton J.A., McInerney J.O. 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. Mol. Biol. Evol. 24:1752–1760.

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

Sievert S.M., Scott K.M., Klotz M.G., Chain P.S.G., Hauser L.J., Hemp J., Hügler M., Land M., Lapidus A., Larimer F.W., Lucas S., Malfatti S.A., Meyer F., Paulsen I.T., Ren Q.,

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Simon J., USF Genomics Class. 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 K.M., El-Mabrouk N. 2012. Gene trees and species trees: irreconcilable differences. BMC Bioinformatics. 13 Suppl 19:S15.

Szöllosi G.J., Tannier E., Daubin V., Boussau B. 2015. The inference of gene trees with species trees. Systematic Biol. 64:e42–62.

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

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

Velasco R., Zharkikh A., Affourtit J., Dhingra A., Cestaro A., Kalyanaraman A., Fontana P., Bhatnagar S.K., Troggio M., Pruss D., Salvi S., Pindo M., Baldi P., Castelletti S., Cavaiuolo M., Coppola G., Costa F., Cova V., Dal Ri A., Goremykin V., Komjanc M., Longhi S., Magnago P., Malacarne G., Malnoy M., Micheletti D., Moretto M., Perazzolli M., Si-Ammour A., Vezzulli S., Zini E., Eldredge G., Fitzgerald L.M., Gutin N., Lanchbury J., Macalma T., Mitchell J.T., Reid J., Wardell B., Kodira C., Chen Z.,

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Desany B., Niazi F., Palmer M., Koepke T., Jiwan D., Schaeffer S., Krishnan V., Wu C., Chu V.T., King S.T., Vick J., Tao Q., Mraz A., Stormo A., Stormo K., Bogden R., Ederle D., Stella A., Vecchietti A., Kater M.M., Masiero S., Lasserre P., Lespinasse Y., Allan A.C., Bus V., Chagné D., Crowhurst R.N., Gleave A.P., Lavezzo E., Fawcett J.A., Proost S., Rouzé P., Sterck L., Toppo S., Lazzari B., Hellens R.P., Durel C.-E., Gutin A., Bumgarner R.E., Gardiner S.E., Skolnick M., Egholm M., Van de Peer Y., Salamini F., Viola R. 2010. The genome of the domesticated apple (Malus × domestica Borkh.). Nature Genetics. 42:833–839.

Waite D.W., Vanwonterghem I., Rinke C., Parks D.H., Zhang Y., Takai K., Sievert S.M., Simon J., Campbell B.J., Hanson T.E., Woyke T., Klotz M.G., Hugenholtz P. 2017. Comparative Genomic Analysis of the Class Epsilonproteobacteria and Proposed Reclassification to Epsilonbacteraeota (phyl. nov.). Front. Microbiol. 8:4962–19.

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

Whidden C., Zeh N., Beiko R.G. 2014. Supertrees Based on the Subtree Prune-and-Regraft Distance. Systematic Biol. 63:566–581.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/758581; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Zhang Y., Sievert S.M. 2014. Pan-genome analyses identify lineage- and niche-specific markers of evolution and adaptation in Epsilonproteobacteria. Front. Microbiol. 5:110.

Acknowledgments

We thank Maxime Godfroid for fruitful discussions. The study was supported by CAPES (Coordination for the Improvement of Higher Education Personnel–Brazil) (awarded to FDKT) and the European Research Council (Grant No. 281357 awarded to TD.)