UNIVERSITY OF CALIFORNIA SAN DIEGO

Efficient Non-parametric Methods for Phylogenomics Inference Using Tree Topology and Branch Length

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Computer Science

by

Uyen Mai

Committee in charge:

Professor Siavash Mirarab, Chair
Professor Vineet Bafna, Co-Chair
Professor Robin Knight
Professor Pavel A. Pevzner
Professor Joel O. Wertheim

2022

The dissertation of Uyen Mai is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

I dedicate this dissertation to my father and mother for their unconditional love and limitless support in my whole life.

To my fiancee, now husband, for always being by my side, showering me with love, and teaching me what true love can do.

To my dear brother and sister-in-law for taking the role of our parents to nurture me, both intellectually and emotionally, in a foreign country.

To my uncles, aunts, cousins, and friends in the US who welcomed me as a family member and had never given up on me.

And to all my teachers and advisors who have taught me invaluable lessons, in and outside the classroom, to help me decide the right career path.

EPIGRAPH

If they give you ruled paper, write the other way

–Juan Ramón Jiménez

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank all the people who have helped me throughout my Ph.D. I express my appreciation to my committee and advisors. I am grateful for the computational support from the San Diego Supercomputer Center (SDSC) through XSEDE allocations and the financial support from NSF, the Microbial Sciences Graduate Research Fellowship, and the Marye Anne Fox Endowed Fellowship for dissertation year.

I enjoy having the support from the excellent lab mates, Erfan Sayyari, Niema Moshiri, Metin Balaban, Maryam Rabiee, Chao Zhang, Eleonora Rachtman, Yueyu Jiang, Puoya Tabaghi, and Eduardo Charvel. Special thanks to Erfan for helping me set up the servers and simulations back in my very first year in the Ph.D. training. His patience helps me gain much confidence in doing independent research in the later years.

I thank all my collaborators in the Web-of-life project. I found myself learning a lot after every single meeting. I give special thanks to Qiyun Zhu and Wayne Pfeiffer, who I had the opportunity to work closely with throughout the years. I learned from them valuable knowledge in microbiology, data analysis, high performance computing, and importantly, the attitude in conducting research. I thank my other collaborators at the Bigelow Laboratory for Ocean Sciences, especially Professor Ramunas Stepanauskas and Julia Brown, for the hours of intellectually stimulating discussions.

I cannot thank Professor Siavash Mirarab enough for his dedication as my Ph.D. advisor. I appreciate his time and efforts for idea brainstorming, problem solving, and paper revision. I am also grateful for his advice on career development and support for every step I put forward. And I thank him for being patient and helping me overcome the stressful time I had during the Ph.D. years.

Chapter 1, in full, is a reprint of the material as it appears in “TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees.” Mai, Uyen, and Siavash Mirarab. BMC genomics 19, no. 5 (2018): 23-40. The dissertation author was the primary

investigator and first author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in “Completing Gene Trees Without Species Trees in Sub-quadratic Time”. Mai, Uyen, and Siavash Mirarab. Bioinformatics (2022). The dissertation author was the primary investigator and first author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in “Minimum variance rooting of phylogenetic trees and implications for species tree reconstruction”. Mai, Uyen, Erfan Sayyari, and Siavash Mirarab. PloS one 12, no. 8 (2017): e0182238. The dissertation author was the primary investigator and first author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in “Log Transformation Improves Dating of Phylogenies”. Mai, Uyen, and Siavash Mirarab. Molecular biology and evolution 38.3 (2021): 1151-1167. The dissertation author was the primary investigator and first author of this paper.

Chapter 6, in full, is a reprint of the material as it appears in “Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea”. Zhu, Qiyun, Uyen Mai, Wayne Pfeiffer, Stefan Janssen, Francesco Asnicar, Jon G. Sanders, Pedro Belda-Ferre, Gabriel A. Al-Ghalith, Evguenia Kopylova, Daniel McDonald, Tomasz Kosciolek, John Yin, Shi Huang, Nimaichand Salam, Jian-yu Jiao, Zijun Wu, Zhenjiang Z. Xu, and Kalen Cantrell, Yimeng Yang, Erfan Sayyari, Maryam Rabiee, James T Morton, Sheila Podell, Dan Knights, Wen-jun Li, Curtis Huttenhower, Nicola Segata, Larry Smarr, Siavash Mirarab, Rob Knight. Nature communications 10, no. 1 (2019): 1-14. The dissertation author was a primary investigator and co-first author of this paper.

VITA

2016	B.S. in Computer Science, Portland State University
2019	M.S. in Computer Science, University of California San Diego
2016–2022	Research Assistant, University of California, San Diego
2022	Ph. D. in Computer Science, University of California San Diego

PUBLICATIONS

Mai, Uyen, and Siavash Mirarab. “Completing Gene Trees Without Species Trees in Subquadratic Time.” Bioinformatics (2022).

Mai, Uyen, and Siavash Mirarab. “Log Transformation Improves Dating of Phylogenies.” Molecular biology and evolution 38.3 (2021): 1151-1167.

Sahoo, Debashis, Livia S. Zaramela, Gilberto E. Hernandez, **Uyen Mai**, Sahar Taheri, Dharanidhar Dang, Ashley N. Stouch et al. “Transcriptional profiling of lung macrophages identifies a predictive signature for inflammatory lung disease in preterm infants.” Communications biology 3, no. 1 (2020): 1-12.

Zhu, Qiyun, **Uyen Mai**, Wayne Pfeiffer, Stefan Janssen, Francesco Asnicar, Jon G. Sanders, Pedro Belda-Ferre et al. “Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea.” Nature communications 10, no. 1 (2019): 1-14.

Balaban, Metin, Niema Moshiri, **Uyen Mai**, Xingfan Jia, and Siavash Mirarab. “TreeCluster: Clustering biological sequences using phylogenetic trees.” PloS one 14, no. 8 (2019): e0221068.

Mai, Uyen, and Siavash Mirarab. “TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees.” BMC genomics 19, no. 5 (2018): 23-40.

Mai, Uyen, Erfan Sayyari, and Siavash Mirarab. “Minimum variance rooting of phylogenetic trees and implications for species tree reconstruction.” PloS one 12, no. 8 (2017): e0182238.

ABSTRACT OF THE DISSERTATION

Efficient Non-parametric Methods for Phylogenomics Inference Using Tree Topology and Branch Length

by

Uyen Mai

Doctor of Philosophy in Computer Science

University of California San Diego, 2022

Professor Siavash Mirarab, Chair
Professor Vineet Bafna, Co-Chair

Computational phylogenetic has been dominated by model-based inference, which is sensitive to model misspecification and has high computational burden. In this dissertation, I present efficient non-parametric methods to facilitate phylogenetic analyses. These methods connect tree topology with the molecular clock principle - as manifested by tree branch lengths - using discrete and numerical optimization. I conclude with an application on one of the largest phylogenomic analyses performed to date.

Chapter 1 presents TreeShrink, an algorithm to detect errors from phylogenetic trees without assumptions about root placement or branch length distributions. I propose and solve

an optimization problem using dynamic programming. By solving this problem, TreeShrink computes the impact of each species on the tree diameter and transforms anomaly detection in tree space to outlier detection in Euclidean space. Our results show effective error detection on simulated and empirical data.

Chapter 2 introduces tripVote, a new method to complete gene trees without a reference species tree. Each gene tree is imputed such that its total quartet distance to the other gene trees is minimized. I develop a quasi-linear time algorithm to solve this problem and show that tripVote is accurate on both simulated and empirical data.

Chapter 3 presents MinVar, a method for tree rooting. MinVar roots a given tree at the point that minimizes the root-to-tip variance. I present a linear-time algorithm to find the MinVar point and prove important properties that make it consistent with the molecular clock principle. Empirically, MinVar is more accurate than other linear-time rooting methods.

Chapter 4 introduces wLogDate, a method that computes divergent times by minimizing the weighted least squares of the log-transformed mutation rates from their mean value. The advantages of this log-transformed penalty function over those derived from a strict-clock model are demonstrated both theoretically and empirically. On simulated data, wLogDate is more accurate than the alternatives in most model conditions and is more than ten times faster than a state-of-the-art method that uses Bayesian MCMC.

Chapter 5 introduces MD-Cat, a method that uses a categorical distribution to approximate the unknown clock model. An EM-based algorithm is described to co-estimate the rate categories and branch lengths in time units. On simulated data of Angiosperm and HIV, the method is more accurate than the alternatives - including wLogDate and Bayesian MCMC - when there are local clocks or heterogeneous rates.

The last Chapter describes an empirical analysis that infers a phylogenetic tree of 10,575 microbial species. This analysis used some of the aforementioned methods and motivated the development of others. The resulting dataset has been used as a reference library in many analyses.

Introduction

Evolution is the fundamental source of the diversity of life on earth [64]. Understanding the evolutionary past is crucial in multiple disciplines in biomedical sciences, including but not limited to microbiology, immunology, vaccine development, epidemiology, and the study of cancer. The evolutionary history of a group of biological entities is referred to as their *phylogeny*, typically represented by a tree structure - a *phylogenetic tree*. The topology of this tree shows how the entities evolved and diverged from their ancestors while the length of each branch quantifies the level of divergence between the two entities (i.e. nodes) that it connects. The divergence can be quantified in different units, such as the expected number of substitutions or the amount of time. Both aspects, topology and branch length, are integral to the interpretation and downstream applications of phylogenetic trees. A main goal of my research has been developing computational methods that consider both aspects.

In the rest of the chapter, I first give a high level overview of the concept used throughout the chapter. I then briefly summarize the contributions from the other chapters.

Background

Phylogenetic inference *Phylogenetic inference* is the process of inferring a phylogenetic tree that best describes the observed characteristics or molecular sequences of extant species (or other biological entities) [130, 94, 188]. Historically, morphology (e.g., color, shape, pattern, size) were used to reconstruct phylogenies. However, modern biology – following the surge in the availability of molecular sequence data in the past several decades [213] – has compelled most researchers to use molecular data (e.g., DNA, RNA, or protein sequences) to infer phylogenies [130].

Maximum likelihood (ML) methods to infer phylogenies from single gene sequences have been used since the mid 20th century [91]. Typically, these methods assume a time-reversible model (for example, the GTR model) and use sequence data to infer the phylogeny that has the maximum likelihood. Although ML-based methods are currently the most commonly used methods for phylogenetic inference, they have limitations, including sensitivity to model misspecification and high computational burden. While model selection techniques (such as [89, 260, 161]) have been proposed to alleviate the former problem, they exacerbate the later. Moreover, scalable ML methods often rely on a time-reversible model - for mathematical and algorithmic convenience. These methods can only infer *unrooted* trees [91] and leave the *problem of root placement* unsolved.

Phylogenomics After the introduction of Next Generation Sequencing [213, 31], using whole genomes to infer phylogenies has become routine, leading to the creation of a new sub-field, called *phylogenomics*. Compared to the traditional phylogenetic inference on a single gene, phylogenomics analyses have access to a collection of genes to infer the species-level phylogeny. While the richer data has a prospect to reduce uncertainties in phylogeny reconstruction [107], the opportunities also come with new challenges [151, 255]. A major challenge in phylogenomics is the biological inconsistencies in the evolutionary history among different parts of the genome (i.e. gene tree) and between a gene tree and the species-level history (i.e. the species tree). To address this problem, new models and methods devoted to phylogenomics inference have been introduced [83, 103, 68].

Molecular clock The phylogenetic trees inferred from sequence data can only have branch lengths in substitution unit. However, the divergence times of the species are of interest, and many downstream applications require or benefit from access to a dated tree (i.e. branch lengths in time unit) or scaled tree (i.e. the unit-height tree proportional to the dated tree). To date or scale the tree, one needs to have a model for mutation rate variation. Following the terminology by Ho *et. al.* [135], I use the term *molecular clock model*, or *clock model* for short, to refer to a model that explains the variation of mutation rates in a phylogeny. Despite its crucial role in many phylogenetic analyses, there is no universally accepted clock model. The true distribution of the mutation rates is unknown and seems to depend on the dataset. Besides, the impact of using an incorrect clock model (or clock model prior) in phylogenetic inference has been studied, but remains controversial [185, 135, 351, 295]. Nevertheless, many methods for tree rooting [44, 342, 340] and dating (see e.g. [135] for a review) rely on a molecular clock model to apply either a maximum likelihood or Bayesian MCMC inference framework.

Several models have been proposed to serve as a molecular clock. The simplest one is the *strict clock* model, which directly applies the original hypothesis by Zuckerkandl and Pauling in 1962 [384] to assume that the substitution rate is constant across all tree branches. Despite its convenience for computation and algorithm development, this model has been shown to be over-simplistic [39, 170]. More sophisticated models to allow rate variation are termed *relaxed clock models*. Relaxed clock models can be classified into two main types: *uncorrelated* and *autocorrelated* models. In a uncorrelated model, the rate of each branch is assumed to be drawn independently from a common underlying distribution. In autocorrelated models, the mutation rate of each branch varies from that of its parent branch under a presumed model. Both types of models usually use a conventional parametric distribution, such as an exponential, gamma, or lognormal distribution, to represent the rate variation [15].

There are also efforts to model the sudden rate shift in evolutionary history, where a lineage possesses a dramatic rate change comparing to its parent and passes on this new rate to its descendants, creating an entire clade or subtree with dramatic rate shift. Such a phenomenon can happen multiple times in the evolutionary history and when a phylogeny possesses this type of rate shift, it is said to have *multiple local clocks* or *heterogeneous rates*. Computational methods addressing heterogeneous rates use a discrete clock model where a finite number of rate changes is assumed in any given tree [373, 370, 14, 79, 101, 126]. With this setting, these methods define local clock as a monophyletic group where every lineage evolves at exactly the same mutation rate. However, empirical studies have explored phylogenies with more complex heterogeneous clock models, such as a mixture of Lognormal distributions [23].

Contributions

Dominated by ML and Bayesian MCMC methods, computational phylogenetic has been driven by model-based parametric inference. These methods are known to be sensitive to model misspecification and tend to have high computational burden. In this dissertation, I present non-parametric and efficient methods to enable several types of phylogenetic analyses, including anomaly detection, data imputation, tree rooting, and tree dating. The first two chapters focus on the issue errors and incompleteness in input data, using branch lengths and the topology, respectively. I follow that with three chapters that are focused on non-parametric models of clock rate, addressing the inter-related problems of dating and rooting phylogenetic trees. I end with an empirical analysis that used some of these techniques and proposed the others.

In Chapter 1, I present TreeShrink, a non-parametric method to detect errors in phylogenetic trees. When they are present, the erroneous sequences often form abnormally long branches and inflate the tree diameter. I define and develop an efficient algorithm to exactly solve the following optimization problem: given an integer k, find the set of k species that should be removed from a tree to minimize its diameter. Using the solutions to this problem with varying k, TreeShrink computes distributions on the impact of individual species on the tree diameter. This way, it transforms the anomaly detection in the tree space to the outlier detection in Euclidean space, which can be solved using standard methods. In addition, TreeShrink can use phylogenomics data to simultaneously filter error from a set of gene trees with high accuracy. When a set of gene trees is given, TreeShrink computes the impact of each species on each gene tree diameter and learns a distribution of this value per species. The outlier detection is then conducted per species to detect all the erroneous sequences that should be removed from the gene tree collection. Our results on simulated and empirical data show fast and effective outlier detection without making explicit assumptions about root placement, branch length distributions, or molecular clock model.

In Chapter 2, I introduce tripVote, a polynomial time algorithm to complete a set of gene trees without a reference species tree. This method works on phylogenomic datasets, which are known to be riddled with missing data: gene trees often lack representatives from some species. Many downstream applications, however, require or benefit from having complete gene trees. While gene tree completion with respect to a reference species tree has been studied before, reference-free completion has not been sufficiently explored. In tripVote, I formulate an optimization problem to complete each gene tree such that its total quartet distance to the other gene trees is minimized. I then design and implement a quasi-linear time algorithm to solve this problem. Interestingly, the same technique can also be used for rooting a gene tree with respect to other gene trees, presenting another application for the technique. Tested on simulated and empirical data for the data imputation task, I show that tripVote is relatively accurate and, unlike reference-based methods for gene tree completion, is unbiased. For rooting trees, comparison to other methods, including those presented in Chapter 3 are more mixed and depend on the model condition.

As mentioned above, phylogenies are often inferred as unrooted trees with branch lengths in substitution units, but many biological applications of the phylogeny, such as viral epidemiology, phylodynamics, and biogeography, require rooted trees with branch lengths in time units. At the same time, the difficulties in clock model selection can reduce the effectiveness of parametric methods for tree rooting and dating, calling for nonparametric alternatives. In the next three chapters, I present nonparametric methods to root and date trees using their branch lengths in addition to topology. In Chapter 3, I present MinVar, a rooting method that roots a given tree at the point that minimizes the variance of its root-to-tip distances. In contrast to the method presented in Chapter 2, this method takes advantage of both topology and branch length. I derive special properties of the MinVar point and show that this rooting method is consistent with the molecular clock hypothesis under the Random Deviations model that I define. I also describe a linear-time algorithm to find the MinVar point and show that it is very scalable in practice. In simulation, MinVar rooting was shown to be more accurate than other linear-time rooting methods.

In Chapter 4 and Chapter 5, I present two novel tree dating methods. Chapter 4 describes wLogDate, which formulates dating as a constrained optimization problem, where the constraints are defined by the calibration points and the objective is to minimize the weighted least squares of the log-transformed mutation rates from their mean value. On a simulated dataset of the HIV envelope gene, wLogDate is more accurate than parametric methods that assume a strict molecular clock. Compared to BEAST - the state-of-the-art method using Bayesian MCMC - wLogDate is more accurate on inter-host data and is more than ten times faster.

Chapter 5 describes MD-Cat, a method that uses a categorical distribution of k bins to approximate the unknown rate distribution. The proposed categorical model is free of assumptions about the true clock model and has a sole user-defined parameter k that determines the resolution of the discretization. I develop an EM-based algorithm to optimize the likelihood function and co-estimate all the k rate categories and tree branch lengths in time units. I test MD-Cat on simulated datasets of Angiosperms and HIV using a wide selection of clock models. Compared to the alternatives, including wLogDate and BEAST, MD-Cat is more accurate on datasets that have local clocks or heterogeneous rates.

I conclude by presenting empirical analyses of a phylogenomics data consisting 10,575 microbial genomes. These analyses used some of the methods described in the other chapters and motivated the development of others. The resulting dataset has been used as a reference library in many downstream analyses.

Chapter 1

TreeShrink: Fast and accurate detection of outlier long branches in collections of phylogenetic trees

Sequence data used in reconstructing phylogenetic trees may include various sources of error. Typically errors are detected at the sequence level, but when missed, the erroneous sequences often appear as unexpectedly long branches in the inferred phylogeny. We propose an automatic method to detect such errors. We build a phylogeny including all the data then detect sequences that artificially inflate the tree diameter. We formulate an optimization problem, called the *k*-shrink problem, that seeks to find *k* leaves that could be removed to maximally reduce the tree diameter. We present an algorithm to find the exact solution for this problem in polynomial time. We then use several statistical tests to find outlier species that have an unexpectedly high impact on the tree diameter. These tests can use a single tree or a set of related gene trees and can also adjust to species-specific patterns of branch length. The resulting method is called TreeShrink. We test our method on six phylogenomic biological datasets and an HIV dataset and show that the method successfully detects and removes long branches. TreeShrink removes sequences more conservatively than rogue taxon removal and often reduces gene tree discordance more than rogue taxon removal once the amount of filtering is controlled. TreeShrink is an effective method for detecting sequences that lead to unrealistically long branch lengths in phylogenetic trees. The tool is publicly available at `https://github.com/uym2/TreeShrink`.

1.1 Background

Datasets used in phylogenetic analyses include a large number of genes and species these days. The number of loci involved and the size of the trees make it impossible to carefully examine every sequence alignment and every gene tree manually. Such manual curation, even if possible, is subject to biases of the curator and poses challenges in reproducibility. But the need for data curation is as strong as ever. Phylogenetic analyses typically use pipelines of many steps, starting from sequencing, to contamination removal, homology and orthology detection, multiple sequence alignment, and gene tree inference, and finally species tree reconstruction. Each step is error-prone, and it has been long recognized that errors can propagate among these steps [36, 145, 156]. However, detecting errors is difficult, especially when large numbers of genes are being analyzed [385]. For example, discordance among estimated gene trees may have biological causes or may be the result of gene tree estimation error; when error-prone gene trees are fed to a species tree estimation method, the error may propagate [182, 216, 106]. This possibility motivates the co-estimation methods that aim to break or weaken the chain of error propagation [16, 5, 328]. However, the end-to-end co-estimation of all steps in the phylogenetic analyses remains elusive [328]. In practice, analysts often devise creative (if ad-hoc) methods to find and remove erroneous data. Such data filtering should be treated with care because it may remove useful signal in addition to error [333], and it also runs the risk of introducing biases. One common method of data filtering is alignment masking [51, 48], despite some criticism [333]. Beyond filtering based on sequences, detecting problematic species from reconstructed trees is also possible.

Two common approaches for filtering based on phylogenetic trees are rogue taxon removal (RTR) [169, 352, 3, 112, 248] and gene tree filtering [139, 323]. More recent approaches include filtering of individual sites with an outsized impact on the tree topology [305]. RTR aim to find species that have an unstable position in the inferred trees, judging the stability with regards to replicate trees generated by bootstrapping [248, 3] or jackknifing [169]. A second approach is to remove genes that are believed to be problematic, perhaps due to missing data [139, 323], lack of signal [287], or even inconsistent signal [305]. When potentially problematic genes are removed, the justification is that the inference of the species tree (i.e., by summarizing gene trees or concatenation) may become more accurate as a result. Alternatively, some analyses (e.g., [355, 298]) filter individual species from individual gene trees based on some criteria (e.g., fragmentation) while keeping the gene. These analyses aim to find and eliminate only the problematic data but nothing more.

The branch lengths of an inferred phylogeny can provide indications of error in sequence data in some cases. If the evolution follows a strict molecular clock, we expect that all leaves should be equidistant from the root. Deviations from the strict clock, if not extreme, would not produce situations where a small minority of species have *dramatically* different rates of evolution and hence root to tip distances. In other words, variations in root to tip distance are expected, but outlier species in terms of distances to the root have to be treated with suspicion. Several types of error in a phylogenetic data, e.g., contamination, mistaken orthology, and misalignment, can lead to the addition of very long branches to the tree (e.g., Figure 1.1a). When a handful of species dramatically diverge from the rest, it is likely that the sequences of outlier species contain errors (of unclear nature). Also suspicious is a species that has normal root to tip distances in most gene trees but has an unexpectedly large root to tip distance in a handful of genes. Even when the sequences of long branches are error-free, they may still pose difficulties due to long branch attraction [24]. Thus, several studies have tried removing species with outlier root-to-tip distances in gene trees [355, 123]. However, rooting is often challenging and prone to error [201]. Moreover, rooting is not necessary for finding outlier species in terms of branch length.

A useful concept is the tree diameter, which gives the maximum distance between any two leaves of the tree. We introduce an optimization problem that if solved efficiently can help in finding species that artificially inflate the tree diameter.

The *k*-shrink problem: Given a tree on n leaves with branch lengths and a number $1 \leq k \leq n$; for every $1 \leq i \leq k$, find a set of i leaves that should be removed to reduce the tree diameter maximally.

We develop an algorithm to solve the problem in $O(k^2h+n)$ time where h is the height of the tree after being rerooted at the centroid edge (which can be done in linear time [201]). Given the solution to the k-shrink problem, we need to decide the species to remove such that the number of error-free sequences removed is minimized. Towards this goal, we propose three statistical tests to find outlier species. We set $k = \Theta(\sqrt{n})$ and compute the proportional reduction in the diameter when going from $i-1$ to i removals for $1 \leq i \leq k$. We then look for outlier values in the spectrum of these proportional reductions; outliers are defined as those that lie at the extreme tails of the distribution, and the outlier detection is controlled by a level of false positive tolerance (α). A further complication is that outgroups, even when error-free, can greatly impact the diameter (Figure 1.1b). Moreover, if a clade has an increased rate of mutations, it may impact the tree diameter more than other clades and may become prone to removal. When multiple gene trees are available, we can learn such patterns of rate variation. Our second statistical test simply combines data from all gene trees to find outliers in a single distribution. The third test goes further and learns a different distribution per species. We implement these tests in a tool called TreeShrink.

We test TreeShrink on six phylogenomic datasets and an HIV transmission tree. We show that TreeShrink improves the quality of gene trees effectively for phylogenomic datasets and can separate strains of HIV. When distributions are learned per species, outgroups are also handled effectively.

1.2 Methods

1.2.1 Notations and definitions

For an unrooted tree t on the leaf-set $\mathcal{L}$, let $\delta(a,b)$ give the distance between a and b. The restriction of t to the leafset A is denoted by $t\restriction_A$, and we use the shorthand $t\backslash_a = t\restriction_{\mathcal{L}-\{a\}}$. We refer to a pair of leaves in t with the highest pairwise distance as a *diameter pair* and call the two leaves *on-diameter*. Any tree has at least one diameter-pair but could have more. We define $\mathcal{P}(t)$ as a set of all diameter pairs of t; that is, $\mathcal{P}(t) = \{(a,b) : (\forall x,y \in \mathcal{L})[\delta(a,b) \geq \delta(x,y)]\}$. The diameter set $\mathcal{D}(t)$ is defined as the set of all on-diameter leaves: $\mathcal{D}(t) = \{a : (\exists x)[(a,x) \in \mathcal{P}(t)]\}$.

We call a tree t *singly paired* if *all* the restricted trees of t (including t) have only one diameter pair; that is, $\forall A \subseteq \mathcal{L}, |\mathcal{P}(t\restriction_A)| = 1$. We refer to the process of removing one leaf from t as a *removal*. A removal is called *reasonable* iff $a \in \mathcal{D}(t)$.

A *removing chain* of t is defined as an ordered list of removals. We refer to a removing chain of length k as a *k-removing chain* and denote it by $\mathcal{H}_k(t)$. We refer to a removing chain that consists only of reasonable removals as a *reasonable removing chain*. An *optimal k-removing chain*, $\mathcal{H}_k^*(t)$, is a removing chain that results in a tree with the minimum diameter among all chains of length k. Any $\mathcal{R}_k(t) \subset \mathcal{L}$ with $|\mathcal{R}_k(t)| = k$ is called a *k-removing set* of t, and is called a *reasonable k-removing set* if there exists an ordering of $\mathcal{R}_k(t)$ that gives a reasonable k-removing chain. We refer to the set of all reasonable k-removing sets as the *k-removing space* of t, and denote it by $\mathcal{S}_k(t)$. We let $\mathcal{R}_k^*(t)$ denote an arbitrary *removing set* that gives the restricted tree with the minimum diameter. Finally, for a rooted version of t, we let $Cld(u)$ denote the set of leaves descended from u.

For all the theoretical results given below, proofs are given in Appendix A.

1.2.2 A polynomial time solution to the *k*-shrink problem

Only reasonable removals have the potential to reduce the tree diameter. If t is singly paired, two reasonable removals exist, and one of them *may* reduce the diameter more. This can be simply checked and thus, the problem is trivial for $k = 1$. For $k > 1$, a greedy approach that takes the optimal removal at each step does not always produce an optimal solution (see Figure A.1a for a counter-example). Therefore, to solve this problem, we need to consider a search space. However, a brute force search for all reasonable k-removing chains is infeasible. The brute force method would first consider the initial diameter pair(s); then, to remove each of the two on-diameter leaves, it would consider the new diameter pair(s) after the first removal and recurse on each diameter-pair. This recursive process produces all reasonable removing chains from 1 to k, but its space grows exponentially.

Three observations enable us to find the optimal solution in a reduced search space that only grows linearly with k. The first observation is that if (a,b) is a diameter pair, then b remains on-diameter after removing a.

Proposition 1. *If an on-diameter leaf is removed, the rest of the on-diameter set are on-diameter for the restricted tree:* $a \in \mathcal{D}(t) \Rightarrow \mathcal{D}(t) - \{a\} \subset \mathcal{D}(t\backslash_a)$.

All i-removing spaces for $1 \leq i \leq k$ can be represented as a directed acyclic graph (Figure 1.2). In this DAG, each node at row i represents an i-removing set $\mathcal{R}_i(t)$, and is also annotated with a diameter pair after the removal of $\mathcal{R}_i(t)$. All the entries in the row i form the $i-$removing space. Any path from the root ending at a node $\mathcal{R}_i(t)$ is an i-removing chain. Note that each node can be reached with multiple paths from the root; this leads to a second observation, which is trivial but important. Any ordering of an i-removing chain gives the same restricted tree. Thus, we can reduce the search space from reasonable chains to reasonable sets. The first two observations allow us to design a polynomial time algorithm for singly paired trees (described next). Our third observation (formalized later) is that when a tree is not singly paired, breaking ties arbitrarily still guarantees optimality.

Singly paired trees

Our main result states that, for singly paired trees, the i^{th} row of the reasonable search space graph (Figure 1.2) contains $i+1$ nodes and one of the nodes gives an optimal i-removing set. Moreover, traversing all $O(k^2)$ nodes in this graph gives the optimal solution to our k-shrink problem.

Theorem 1. *The k-removing space of a singly paired tree t includes all the optimal k-removing sets of t; that is:* $\forall k > 0 : \mathcal{R}_k^*(t) \in \mathcal{S}_k(t)$.

Theorem 2. *The size of the k-removing space for a singly paired tree t is* $k+1$.

Corollary 1. *The size of the reasonable search space up to level k is* $O(k^2)$.

Our algorithm (Algorithm 1) start with a preprocessing in order to enable computing the pair set at any point in a removal chain in $O(n)$. The preprocessing uses a bottom-up traversal of t (rooted arbitrarily). For each internal node u, we store four values: (i) the leaf $x \in Cld(u)$ with the longest distance to u, (ii) the distance $\delta(u,x)$, (iii) the leaf $y \in Cld(u) - Cld(c_1)$ with the longest distance to u (where c_1 is a child of u such that $x \in Cld(c_1)$), and (iv) the distance $\delta(x,y)$ (see Figure A.1b). We store these values for each node u as a tuple $rec(u) = (rec_1(u), rec_2(u), rec_3(u), rec_4(u))$. These values can be computed in a post-order traversal of the tree in the natural way. Once these records are computed, finding diameter pairs can be done quickly (see function `FindPair` in Algorithm 1). Let (a,b) be a diameter pair; note that regardless of the arbitrary rooting of the tree, at the LCA of a and b, the record includes a, b as the first and third fields and the tree diameter as the last. Thus, the tree diameter corresponds to the record with the largest fourth value. As we will see, throughout the algorithm, the values of the records may have to change. However, these updates can also happen in $O(h)$. Thus,

Algorithm 1 Polynomial time k-shrink algorithm. Function `Shrink` gives the main algorithm. Assuming $(a,b) \in \mathcal{P}(t)$, function `FindPair` finds a leaf x such that $(x,b) \in \mathcal{P}(t\backslash_a)$; it assumes that t has $rec(u)$ computed for all of its nodes.

```
function SHRINK($t,k$)
    Compute $rec(u)$ for all internal nodes $u$ of $t$ using a postorder tree traversal
    $(a,b) \leftarrow (rec_1(u), rec_3(u))$ where $u$ is the node with the maximum $rec_4(u)$
    $minD \leftarrow$ an array of $k$ elements initialized to $\infty$
    $Q \leftarrow$ an empty queue initialized with tuple $(0,a,b,\{\},\delta(a,b))$
    $seen \leftarrow \emptyset$
    while $|Q| \neq 0$ do
        $(i,a,b,R,d) \leftarrow Q.remove()$
        $minD[i] \leftarrow \min(minD[i],d)$
        if $i < k$ then
            $Q.append(i+1, FindPair(t\upharpoonright_{\mathcal{L}-R},a,b),b,R \cup \{a\})$
            if $i \notin seen$ then
                $Q.append(i+1,a,FindPair(t\upharpoonright_{\mathcal{L}-R},b,a),R \cup \{b\})$
                $seen \leftarrow seen \cup \{i\}$
    return $minD$
function FINDPAIR($t,a,b$)
    $diameter \leftarrow 0$
    for all node $u$ in the path from the parent of $a$ to the root do
        Update $rec(u)$ from records of its children (ignore $a$ if it is one of the children)
        if $rec_4(u) > diameter$ then
            $diameter \leftarrow rec_4(u)$ and $diamPair \leftarrow (rec_1(u), rec_3(u))$
    for all node $u$ in the path from the parent of $b$ to $LCA(a,b)$ do
        if $rec_4(u) > diameter$ then
            $diameter \leftarrow rec_4(u)$ and $diamPair \leftarrow (rec_1(u), rec_3(u))$
    return $x \in diamPair$ if $x \neq a$
```

Proposition 2. *Given a rooted tree t of height h, $(a,b) \in \mathcal{P}(t)$, and $rec(u)$ for all nodes $u \in \mathcal{L}$, we can find one diameter pair of $t\backslash_a$ in $O(h)$.*

Once the preprocessing finishes, we start building the DAG (Fig 1.2). We start with the root node, corresponding to the initial tree t and build the rows iteratively. For any node at step i with (x,y) as its diameter pair, two nodes have to be added to the next row, one for removing x and another for removing y. As the proof of Theorem 2 (Appendix A) indicates, two sister nodes in step i have to share one descendant in step $i+1$ (Figure 1.2). Thus, to construct each row from the previous row we simply need to find the diameter pair of the tree restricted to the removal-set of each node; this is done in the function `FindPair` previously described. As we build the DAG, we also keep track of the length of the diameter at each node and the optimal i-removing set. At the end, we report an optimal-removing set for each i from 1 to k.

According to Proposition 2, finding each new diameter pair after removing any node can be done in $O(h)$. From Corollary 1 and Proposition 2, we have:

Corollary 2. *Algorithm 1 solves the k-shrink problem in* $O(k^2h+n)$.

Generalization to all trees

If the tree t is not singly paired, nodes in the search graph could have more than two children which increase the size of the search space. However, we prove that we can break the ties arbitrarily at any step and still guarantee *an* optimal solution. It follows naturally that Algorithm 1 also works for trees that are not singly paired.

For any diameter pair $(a,b) \in \mathcal{P}(t)$, we define a *pair-restricted k-removing space* as a subset of $\mathcal{S}_k(t)$ such that each of its elements includes either a or b.

Theorem 3. *For any k, any arbitrary pair-restricted k-removing space includes at least one optimal k-removing set.*

Proof (sketch). It is not hard to prove that any tree t has a single midpoint which partitions its diameter set into disjoint subsets. We call each of those subsets a *diameter group* of t (Appendix A, Lemma S2 and Lemma S3). Clearly, unless all but one of the diameter groups are removed, the tree diameter is unchanged. We refer to the restricted tree of t that have all but one of the diameter groups removed as a *minimum shrunk tree* of t. We can prove that any arbitrary pair-restricted removing space can produce *all* minimum shrunk tree (see the full proof in the Appendix A). If k is so small such that there is no k-removing set can reduce the tree diameter, any solution is optimal and the result of Theorem 3 trivially follows. Otherwise, any optimal solution of k-shrink can be induced from one of the minimum shrunk trees (Lemma S4 in Appendix A, Additional file 1). Thus, to find an optimal tree t^*, we can start from *any* pair-restricted removing space and concatenate the two removing chains: the chain that induces the minimum shrunk tree t_i^* from any arbitrary diameter pair, and the chain that starts from t_i^* to induce t^*. Full proof is given in Appendix A. □

According to Theorem 3, any pair-restricted k-removing space includes at least one optimal solution. For a tree that is not singly paired, we can arbitrarily restrict the search space to any of its diameter pairs *at any step* of the algorithm. This ensures that the search space size grows with $O(k^2)$, and that Algorithm 1 still correctly finds an optimal solution in $O(k^2h+n)$.

1.2.3 Statistical selection of the filtering species

The solution to the k-shrink problem for a given k gives the minimum diameters for $1 \leq i \leq k$ and the corresponding optimal removing sets. Given these results, we now need to find a set of species that have *unexpectedly* large impacts on the tree diameter. Defining what is an expected impact on the diameter is not trivial and depends on many factors such as the rate of speciation, taxon sampling, and the tree topology. To avoid modeling such processes, we use empirical statistics.

Let ν_i be the ratio of the minimum diameter with $i-1$ leaves removed and the minimum diameter with i leaves removed, and let $\Delta_i = \log(\nu_i)$. For a tree with no outlier branches, we expect ν_i values to be close to one (e.g., T1 in Figure 1.3a). For a tree with one outlier leaf on a very long branch, we expect that ν_1 is much larger than other ν_i values (T2 in Figure 1.3a). If two species are on a very long branch, we expect a small ν_1, a large ν_2, and small values again for $i > 2$ (T3 in Figure 1.3a). If there are two exceptionally long branches, one with three species and another with five species, we expect ν_3 and ν_8 to be large and other values to be small (T4 in Figure 1.3a). We use ν values to detect outliers, but we first need to introduce the concept of a *signature*.

The ν_i values for the removing sets that include a species measure the impact of that species on the tree diameter. We will refer to the maximum Δ_i among all removing sets i that include a species as the *signature* of that species (note that this is defined only for some of the species). A species with an exceptionally larger signature compared to the other species can be considered an outlier (Figure 1.3b). To define what qualifies as exceptionally large, we design three different tests. The first test can be applied to a single tree, while the other two require a collection of gene trees.

The "per-gene" test

For a single input tree and a large enough k, we have a distribution over signature values. Since we have limited data in this scenario, we use a parametric approach and fit a log-normal distribution to the signatures. Given a false positive tolerance rate α, we define values with a CDF above $1-\alpha$ as outliers. Then, species associated with the outlier signatures are removed.

The "all-gene" test

When a dataset includes several gene trees, all related by a species tree, combining the distributions across genes can increase the power. With many genes, we may also be able to distinguish outgroup species from outliers. The signatures of outgroups across all gene trees should be consistently higher than those of ingroups, and these high signatures will appear as part of the combined signature distribution. Thus, we may be able to avoid designating outgroups signatures as outliers.

In this test, we put the signature of all genes together to create one distribution. Unlike the per-gene test, here we have many data points, which enables us to use a non-parametric approach. We compute a kernel density function [310] over the empirical distribution of the combined set of signature values. To estimate the density, we use Gaussian kernels with Silverman's rule of thumb smoothing bandwidth [310] (as implemented in the R package [337]). Given the density function and a false positive tolerance rate α, we define values with a CDF above $1-\alpha$ as outliers.

The "per-species" test

Outgroups can contribute to the tree diameter as much as erroneous species (Figure 1.1b). To better distinguish outgroups from errors, when a set of gene trees are available, we can learn a distribution *per* species. Given a sufficient number of gene trees, the signatures of a species across all genes form a distribution that specifically captures the impact of that species on the gene tree diameter. These species-specific distributions naturally model the inherent difference between outgroups and ingroups in terms of their impacts on the tree diameter. More broadly, changes in the evolutionary tempo are captured naturally by the per species distributions.

In this test, we first compute a non-parametric distribution of the signature values for each species. When the signature of a species is not defined for a gene, we simply use zero as its signature. Then, for each species, we use the same non-parametric approach as in the all-gene test to compute a threshold for the signature value corresponding to the chosen α. Finally, we remove each species from those genes where its signature is strictly above its species-specific threshold.

The default parameters of TreeShrink

TreeShrink has two parameters: α and k. By default, we set α to 0.05 (but users can choose other thresholds). Large values of k do not fit our goal of finding *outlier* species and can even lead to misleading results (e.g., Figure A.2), but a small value of k may also miss outliers and may lead to insufficient data points for learning distributions.

Using a value of k that grows sublinearly with n (i.e., the number of leaves) gives us an algorithm that is fast enough for large n. For example, using $k = \Theta(\sqrt{n})$ gives $O(nh)$ running time, which on average is close to $O(n \log n)$ and is $O(n^2)$ in the worst case. While the choice must be ultimately made by the user, as a default, we set $k = min(\frac{n}{4}, 5\sqrt{n})$. This heuristic formula ensures that our running time does not grow worse than quadratically with *n* but also avoids setting *k* to values close to *n* (thus also limits the proportion of leaves that *could* be removed).

1.2.4 Evaluation procedures

Datasets

We use six multi-gene datasets and a single-gene HIV dataset, and each dataset includes one or more outgroup species defined by the original papers (Table 1.1).

Plants [355]: This dataset of 104 plant/algae species (four Chlorophyta outgroups) and 852 genes was used to establish early diversification patterns within land plants and their sister groups. The data are based on transcriptoms, and authors faced challenges in terms of gene identification and annotation, leading to abundant missing data. To obtain reliable species trees using ASTRAL [218], the authors had to use various filterings, including removal of low occupancy genes and genes with fragmentary sequences. The ASTRAL tree obtained on these filtered gene trees was mostly congruent with results from concatenation, though some interesting clades (e.g., the Bryophytes) were differently resolved. In our analyses, we start with all gene trees estimated from nucleotide data with the third codon position removed.

Insects [220]: This phylotranscriptomic dataset includes 144 species and 1,478 genes. This dataset was used to resolve controversial relationships among major insect orders, but only concatenation analyses were reported. A different paper performed a species tree analysis of the same dataset using ASTRAL, obtained on RAxML gene trees that we estimated on all 1,478 gene trees [298]. We use these gene trees in our analysis.

Metazoa-Cannon [47] and Rouse [282]: Whether Xenacoelomorpha (a group of bilaterally symmetrical marine worms) are sister to all the remaining Bilateria (animals with bilateral symmetry) has been the subject of much recent debate [47, 254, 282]. The Cannon *et al.* dataset of 213 genes from 78 species sampled from across the animal tree-of-life was used to confidently place Xenacoelomorpha as sister to Bilateria. Among other analyses, ASTRAL-II [219] was used on a collection of gene trees that the authors published, and we use. The dataset by Rouse *et al.* addresses the same question as Cannon *et al.* using 393 genes and 26 species.

Mammals [314]: This mammalian dataset consists of 37 species (36 mammals and Chicken as outgroup) and 424 gene trees. Since the original gene trees had several issues (including insufficient ML searches and mislabeled species [318]), here we use RAxML gene trees that were inferred and used in a re-analysis of this dataset [216]. Several reanalyses of this dataset using various methodologies have largely been consistent, except, for the position of the tree shrews that often changes [216, 218].

Frogs [96]: This dataset consists of 164 species (156 frog species and 8 outgroups) and 95 genes. The dataset was used to study the evolutionary history and tempo of frog diversification [96]. The RAxML gene trees we use here were used as inputs for ASTRAL to construct the species tree [96] and were provided by the authors [95].

HIV dataset [189]: This HIV dataset consists of 648 partial HIV-1 *pol* sequences that were used to reconstruct the local HIV-1 transmission network from 1996 to 2011 in San Diego, California. The dataset consists of 639 subtype B, 7 non-subtype B, and 2 unassigned sequences of HIV-1 *pol* coding region. The sequences have GenBank accession numbers from KJ722809 to KJ723456, and were provided to us by the authors. Note that this dataset has only one gene.

Methods tested

We implemented TreeShrink (`https://github.com/uym2/TreeShrink`) using the Dendropy package [327]. We compare the three tests of TreeShrink, namely per-gene, all-gene, and per-species. In addition, we compare the most effective test of TreeShrink, the per-species test, with two alternative methods and a control where we remove species randomly from the tree.

The main alternative to TreeShrink used previously [355, 123] is to root gene trees and then remove species with outlier root-to-tip distances. We use this "rooted pruning" approach where we define outliers as values that lie several standard deviations (we vary this threshold) above the average. For the Plant dataset, 681 genes included the outgroups; for the remaining, we used a linear-time implementation of the midpoint rooting [201]. In other datasets, each gene tree included at least one of the outgroups. While the goals of RTR are somewhat different from ours, we also compare our method with RogueNaRok [3], which defines a rogue taxon as one that has unstable positions in replicate bootstrap runs.

Evaluation

Judging the effectiveness of the filtering methods on real data is challenging, as we do not know if a removed sequence is in fact erroneous. However, patterns of discordance can help. While true gene trees may be discordant with the species tree, erroneous sequences will further increase the observed discordance. Thus, the amount of gene tree discordance among genes should reduce as a result of effective filtering, and more effective filtering methods arguably reduce the discordance more than less effective ones. Thus, the quality of a filtering procedure can be judged (albeit with some uncertainty) by its impact on gene tree discordance, as long as its optimization problem does not seek to reduce discordance directly. Note that none of the methods that we test take the species tree as input, and none is trying to directly reduce the gene tree discordance. Thus, we use the reduction in discordance as one measure of accuracy. To compute gene tree discordance, we compare all pairs of gene trees to each other and use the MS (Matching Splits) metric [28], implemented in the TreeCmp [29] to measure distance. To facilitate the interpretation of MS, which is not normalized, we include random removal as a control.

A second concern is the potential of methods to aggressively remove true signal. To evaluate this, we investigate the impact of filtering on the taxon occupancy, defined as the number of gene trees that include each species. Lowered occupancy may negatively impact downstream analyses such as species tree inference and functional analyses. Ideally, a filtering method would not reduce taxon occupancy dramatically. Moreover, removing the same species repeatedly from many genes could be even more problematic for downstream analyses such as species tree estimation.

Filtering methods have a knob that can control the amount of filtering. To avoid impacts of arbitrary choices, we explore a range of knob settings. For the three tests of TreeShrink, we set α to 20 different values in the range $[0.005, 0.1]$. For RogueNaRok, we change the weight factor to control the penalizing factor of the dropset size by setting it to 21 values in range $[0, 1.0]$ (0.0 is the default value). For rooted pruning, we vary the number of standard deviations above the average that would constitute long branches between 0.25 to 5.00, with 0.25 increments. For random pruning, for each threshold of TreeShrink on each gene tree, we remove the same exact number of leaves as TreeShrink removes, but we choose the species randomly. We repeat the random pruning ten times and show the average.

On the HIV dataset, we test the power of TreeShrink ($\alpha = 5\%$), rooted pruning (3 std), and RogueNaRok (default settings) in detecting the outliers. Outliers are either non-subtype B sequences in the full dataset in experiment 1 or the simulated outliers we added in experiment 2 (described below).

In the first experiment, we infer a RAxML tree from the 648 sequences and use it as the input for TreeShrink. We root the RAxML tree at its midpoint and use it for rooted pruning. To run RogueNaRok, we also create 100 bootstrap trees using RAxML. We use the 7 non-subtype B and 2 unassigned sequences as *outliers* (see Table A.1) and test if TreeShrink, rooted pruning, and RogueNaRok can detect them.

In the second experiment, we add 10 simulated outliers to the 639 subtype B sequences and use TreeShrink and rooted pruning to detect them. To create the outliers, we randomly select 10 sequences from the 639 subtype B sequences and change a small fraction of their sites, selected randomly, to a random nucleotide drawn from the distribution of the base frequencies estimated from the original sequences. In order to root the tree, we include the 3 subtype C sequences (Table A.1) and root the tree on the branch separating the two subtypes, then remove them before feeding it to TreeShrink or rooted pruning. We create two sets of data, one with 5% and the other with 10% of the sites changed, each consists of 20 replicates. The trees in this experiment are estimated by FastTree.

1.3 Results

We start by comparing the three tests currently implemented in TreeShrink. We then compare the per-species test of TreeShrink with alternative methods.

1.3.1 Comparing the three tests of TreeShrink

The impact of filtering on taxon occupancy

The three tests of TreeShrink ($\alpha = 0.05$) impact taxon occupancy differently, especially for outgroups. Outgroups naturally impact the tree diameter, but ideally, they should not be removed more often than other leaves. In all six datasets, the per-gene and all-gene tests tend to remove outgroups aggressively, while the per-species test removes all species, including outgroups, close to uniformly (Table 1.2 and Figure A.3).

The most severe case is chicken, the sole outgroup in the Mammalian dataset. Chicken is removed in 12% of the genes by the per-gene test (19 times more than the average) and in 17% by the all-gene test (13 times more than the average). Note that in this dataset, both per-gene and all-gene tests remove only around 1% of the data, so the frequent removal of the chicken corroborates our suspicion that TreeShrink used with per-gene or all-gene tests can remove outgroups often even if the outgroup sequence contains no errors. The per-species test, on the other hand, only removes chicken slightly more often than the average: it removes about 4% of the overall data and removes chicken in about 5% of the genes that have it (Figure A.3b).

In addition to the outgroups, platypus is also removed often. Being basal to the other mammals, platypus is prone to the same issues as outgroups. However, there is also some evidence that platypus is often misplaced in many gene trees of this dataset [318]. Just as the chicken, platypus is removed significantly more often than other species: 5% in the per-gene test (7 times more often than the average) and 13% in the all-gene test (10 times more often than the average). Again, the per-species test removes platypus just slightly more often than the average: platypus is removed in about 5% of the genes while the average of all species is 4% (Figure A.3).

The impact of filtering on gene tree discordance

We now compare the three tests of TreeShrink in reducing gene tree discordance with minimal filtering. A method is preferred when it reduces the discordance more for a given level of filtering (i.e., higher lines in Figure 1.4 are preferred). Except for the Frogs dataset, all the three tests of TreeShrink are on average better than the control random pruning. On the Frogs dataset, however, only the per-species test is better than the control. The failure of the other two tests could be because they remove outgroups often (see Table 1.2) and fail to remove the true outliers (perhaps because the true outliers are masked by the outgroups). Overall, the per-species test is consistently the most effective, followed by the all-gene test, and finally the per-gene test. Differences between the per-species test and the all-gene tests are substantial for plants, mammals, and frogs datasets, and less pronounced for others. Since the per-species test of TreeShrink is consistently the best here, we recommend using the per-species test for phylogenomic datasets which contain many genes.

1.3.2 Comparing TreeShrink per-species with RogueNaRok and rooted pruning

We now compare TreeShrink per-species with alternative filtering methods.

The impact of filtering on taxon occupancy

Methods run in the default settings ($\alpha = 0.05$ for TreeShrink, 3 std for rooted pruning) impact occupancy differently. Overall, RogueNaRok reduces the occupancy more than the other methods (Figure 1.5). Single species at the base of large clades seem especially prone to filtering by RogueNaRok. In contrast, TreeShrink and rooted pruning do not remove any specific taxon extensively.

On the plants dataset (Figure 1.5a), RogueNaRok removes three species from at least half of the gene trees where they are present and removes 12 species from one-third of the genes. Examples include *Kochia scoparia* (removed in 343 out of 654 genes), *Acorus americanus* (251/693), and *Larrea tridentata* (221/590) genes. *Kochia scoparia* is on a long branch and sister to a group of 7 Eudicots, and *Acorus americanus* is basal to 10 Monocots [355]. Surprisingly, *Arabidopsis thaliana* is removed in 200 genes, even though it is a genome and is presumably less error-prone compared to most other transcriptomes species. Moreover, a focal point of this study is placing *Chara vulgaris* as basal to all land plants plus two algal groups (Zygnematophyceae, and Coleochaetales). RogueNaRok removes Chara from 160 genes out of 302 that include it; such aggressive filtering could limit the ability to answer this main biological question with confidence. In contrast, rooted pruning and TreeShrink remove 4% and 7% of the data, respectively. TreeShrink never removes any species in more than 6% of the genes and all species are removed in similar proportions.

On the insects dataset (Figure 1.5c), RogueNaRok removes 17% of all the data and many removed species are basal to large diverse groups. For example, RogueNaRok removes

Conwentzia psociformis, which is basal among 8 Neuropterida [220] from 684 out of 1,412 genes that included it. *Zorotypus caudelli*, an enigmatic species placed as sister to a large clade in the ASTRAL species tree is also removed from 52% of the genes. Interestingly, RogueNaRok removes several outgroups, including *Speleonectes tulumensis* and *Cypridininae sp* frequently (56% and 57%). In contrast, rooted pruning and TreeShrink only remove a minimal amount of data (1% and 4%, respectively) and do not impact occupancy dramatically for any species.

Similar patterns are observed on Metazoa datasets (Figure 1.5ef). RogueNaRok removes more than 20% of the leaves overall, and many species are extensively removed from many genes. In the Cannon dataset, *Xenoturbella bocki* is removed in 93 out of the 208 genes that included it. Xenoturbella is the basal branch of the Xenacoelomorpha and in this study, is one of the most important species; removing it in 45% of genes would leave a long branch and could negatively impact the placement of Xenacoelomorpha. Rooted pruning and TreeShrink, again, remove a minimal portion of the data (2% and 4%, respectively) and no species is extensively removed.

The mammalian dataset is not extensively filtered by any method (Figure 1.5b). Rooted pruning only removes about 1% of the data, while RogueNaRok and TreeShrink remove about 4%. RogueNaRok removes three species (shrew, tree shrew, and hedgehog) relatively often (i.e., $>$ 80 genes). The shrew and the hedgehog are both basal branches to a larger clade of Laurasiatheria. The tree shrew has a very uncertain position in various species trees estimated on this dataset [216, 218, 314, 318]; RogueNaRok results indicate that its position is also unstable in gene trees. Platypus is also removed relatively often by rooted pruning (54 times), but somewhat less frequently by RogueNaRok (31 times) and TreeShrink (20 times). Several issues in the platypus sequences have been identified [318], and perhaps, its extensive filtering by rooted pruning is justified. Similar to the mammalian dataset, on the frogs dataset (Figure 1.5d), all methods remove very little data ($<$3% overall).

The impact of filtering on gene tree discordance

Since extensive filtering is neither intended nor desired in this section, we focus on filtering thresholds that result in removing at most 5% of the data (see Figure A.4 for the full data). On all six datasets, all the three filtering methods are on average better than the control random pruning. Comparing TreeShrink and the two alternatives, different patterns are observed on various datasets (Figure 1.6).

On the two datasets with the largest numbers of genes, Plants and Insects, TreeShrink outperforms the other methods substantially (Figure 1.6ab). On the Insects dataset, RogueNaRok barely outperforms random pruning and TreeShrink is substantially better than rooted pruning. On the Plants dataset, rooted pruning and RogueNaRok are essentially tied and TreeShrink is consistently better than both. For example, TreeShrink with a 0.03 threshold removes 1476 species in total from all genes and reduces the average pairwise MS discordance by 15 units (as opposed to 11 for the control), whereas RogueNaRok and rooted pruning need to remove 1649 and 1740 species to achieve a reduction of up to 15 units in the MS discord.

On the Metazoa-Cannon dataset (Figure 1.6c), TreeShrink and RogueNaRok both outperform rooted pruning, and TreeShrink has a small but consistent advantage over RogueNaRok. On the Metazoa-Rouse dataset, all methods are comparable and barely outperform random pruning (Figure 1.6d).

On the Mammalian dataset (Figure 1.6e), RogueNaRok is by far the best, followed by TreeShrink and rooted pruning, which have similar overall performance. On the Frogs dataset, which included only 95 genes, RogueNaRok and rooted pruning are tied and both substantially outperform TreeShrink (Figure 1.6f).

Overall, TreeShrink is the best or tied with the best method in four datasets, and is outperformed in the other two. TreeShrink seems especially well suited for datasets with a large number of genes.

1.3.3 The HIV dataset

Detecting non-subtype B sequences

Using the RAxML tree of the 648 HIV *pol* sequences as input, TreeShrink correctly detects all seven non-subtype B sequences, including a single subtype CRF01_AE sequence, two CRF02_AG sequences, three subtype C sequences, and a subtype G sequence. The two unassigned sequences are not identified as outliers by TreeShrink (Figure 1.7). Importantly, TreeShrink does not remove any subtype B sequences. In contrast, RogueNaRok identifies 41 rogue sequences in total, only one of which is non-subtype B (the subtype G sequence KJ723366). As we elaborate in the discussions, these differences are due to different objectives of the two methods. With midpoint rooting, rooted pruning detects three non-subtype B sequences (i.e., CRF01_AE and two CRF02_AG) as outliers but it misses the other 4 non-subtype B sequences and has two false positives.

Detecting simulated outliers

Recall that for each simulated dataset, we have 20 replicates and each consists of 10 simulated outliers, for the total of 200 outliers to be detected. On the dataset with outliers at 10% changed in sequences, TreeShrink correctly detects 198/200 outliers and rooted pruning detects all 200/200 outliers; neither method has a false positive. On the dataset with outliers at 5% changed in sequences, TreeShrink correctly detects 106/200 outliers with 9 false positives while rooted pruning detects 131/200 outliers with 17 false positives. Overall, TreeShrink has higher precision and specificity but lower sensitivity comparing to rooted pruning (table 1.3), indicating that TreeShrink is a more conservative approach. Figure 1.8 shows one example for each of the two simulation settings.

1.4 Discussions

It has been noted before that extreme long branches in a phylogeny can be erroneous. Gatesy and Springer used the presence of long branches in gene trees estimated in two mammalian datasets to argue against specific coalescent-based analyses (see Figs. 9 and 10 of their paper [106]). To eliminate problematic long branches, a typical approach is to root the tree and filter out leaves too distant from the root [355, 123]. TreeShrink can automatically filter out such outliers *without* rooting. In addition, TreeShrink is very scalable. It could finish processing the GreenGenes tree [70] with $203,452$ leaves ($k = 2255$) in 28 minutes and identified 39 species that could be filtered.

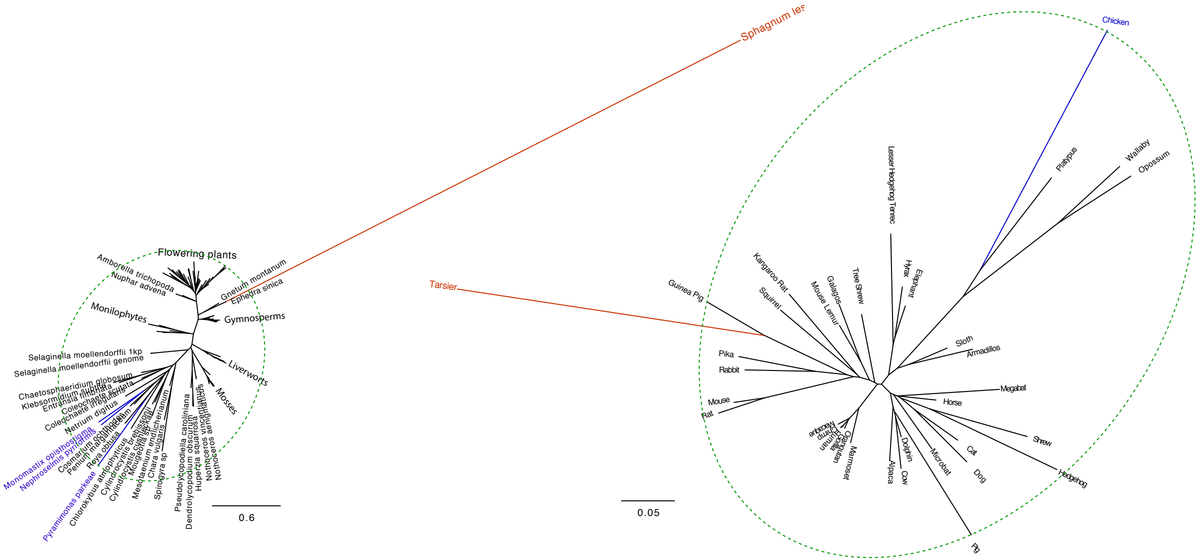
In this study, we observed that the per-species test of TreeShrink is consistently the best strategy, followed by the all-gene and the per-gene tests. However, it should be noted that the per-species test requires more data than the two alternative tests, and its data requirement has some practical implications. Because it relies on computing a distribution per species by aggregating data from all gene trees, the per-species test may degrade in performance when few genes are available. Consistent with this observation, we observed that the only dataset where the per-species test of TreeShrink was outperformed by rooted pruning was the Frogs dataset, which has fewer than a hundred gene trees (less than half of any other datasets). Similarly, the per-species test may not have enough information for species that have extremely low occupancy, to begin with. Therefore, we recommend caution in taking the suggestions of the per-species test for low-occupancy species.

We only examined effects of filtering leaves from existing trees without redoing alignments or gene trees after filtering. This was mostly due to our inability to replicate the exact analysis pipelines of every dataset we analyzed. When used on novel datasets, it is better to reestimate alignments and gene trees after the problematic sequences have been removed, because the problematic sequences could have negatively impacted gene alignments and gene trees of the remaining sequences.

Although we compared our method to RogueNaRok as an alternative to our approach, we point out that the two methods have different objectives and can complement each other. While RogueNaRok aims to remove *rogue species* based on topological stability, TreeShrink detects and removes *erroneous species* based on tree diameter. An analysis pipeline could use a combination of the two methods to find both erroneous sequences and difficult unstable tips of the phylogenetic tree. Our HIV dataset is a case in point. The differences between TreeShrink and RogueNaRok on this dataset can be mainly attributed to their different objectives. TreeShrink is specialized for detecting outlier species and is well-suited for specific applications such as screening of sub-types, finding contamination, or perhaps even finding paralogs. RogueNaRok, on the other hand, is designed to find species with unstable positions. Thus, our results should not discourage the use of RogueNaRok. Rather, the HIV example, and our results more broadly, are meant to clarify that shrinking the tree diameter can be an orthogonal approach to rogue taxon removal.

1.5 Conclusions

In this paper, we introduced TreeShrink, a method to remove species that disproportionately impact a phylogenetic tree diameter *without* rooting. The tool is fully automatic and is publicly available. In our study, we showed that TreeShrink is highly accurate in screening subtypes of HIV, and is effective in reducing gene tree discordance in phylogenomic datasets. As a complement to the state-of-the-art rogue taxon removal tools, TreeShrink can be a new component to an analysis pipeline for screening sub-types, filtering contamination, and detecting paralogs.



a) Easy case: a gene tree in the Plant dataset

b) Hard case: a gene tree in the mammalian dataset

Figure 1.1: Example trees with suspicious long branches.(*a*) An unfiltered gene tree of a Plant dataset [355] with an obvious outlier leaf; (*b*) a gene tree in a mammalian dataset with a hard to detect outlier branch [314]. Outgroups are shown in blue and the suspicious long branches in the red. Dashed green line: the tree diameter after removal of red branches. Detecting the red branch is easy on the left but hard on the right.

1.6 Acknowledgements

Authors thank Joel Wertheim for providing the HIV dataset.

Chapter 1, in full, is a reprint of the material as it appears in “TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees” (2018). Mai, Uyen; Mirarab, Siavash, BMC genomics, 19(5), 23-40. The dissertation author was the primary investigator and first author of this paper.

Dataset	Species	Genes	Outgroups	Download
Plants [355]	104	852	Monomastix opisthostigma, Uronema sp., Nephroselmis pyriformis, Pyramimonas parkeae	DOI `10.1186/2047-217X-3-17`
Mammals [314]	37	424	Chicken	DOI `10.13012/C5BG2KWG`
Insects [220]	144	1478	IXODES SCAPULARIS, Symphylella vulgaris, Glomeris pustulata, Lepeophtheirus salmonis, DAPHNIA PULEX, Cypridininae sp, Sarsinebalia urgorii, Celuca puligator, Litopenaeus vannamei	`http://esayyari.github.io/InsectsData`
Cannon [47]	78	213	Salpingoeca rosetta, Monosiga brevicollis Mnemiopsis leidyi, Pleurobrachia bachei, Euplokamis dunlapae	DOI `10.5061/dryad.493b7`
Rouse [282]	26	393	Mnemiopsis leidyi, Amphimedon queenslandica, Trichoplax adhaerens, Nematostella vectensis	DOI `10.5061/dryad.79dq1`
Frogs [95]	164	95	Latimeria chalumnae, Protopterus annectens, Homo sapiens, Crocodylus siamensis, Gallus gallus, Ichthyophis bannanicus, Batrachuperus yenyuanensis, Andrias davidianus	DOI `10.5061/dryad.12546.2`

Table 1.1: Summary of the biological datasets

Dataset	Method	Portion of data removed(%)	Portion of outgroups removed(%)
Plants	per-gene	3.3	29.9
	all-gene	2.5	12.8
	per-species	4.9	5.1
Mammals	per-gene	0.6	11.8
	all-gene	1.2	17.0
	per-species	3.6	4.7
Cannon	per-gene	1.4	6.2
	all-gene	1.3	4.7
	per-species	3.5	5.0
Rouse	per-gene	1.3	1.9
	all-gene	1.2	1.1
	per-species	4.0	4.5
Insects	per-gene	1.2	6.6
	all-gene	0.8	2.9
	per-species	4.3	5.0
Frogs	per-gene	1.3	26.7
	all-gene	0.8	15.9
	per-species	2.7	4.5

Table 1.2: The impact of the three tests of TreeShrink on taxon occupancy

Dataset	Method	True positives	False positives	Precision	Recall (Sensitivity)	Specificity
5%	TreeShrink	106	9	**92.2%**	53.0%	**98.6%**
	Rooted pruning	131	17	88.5%	**65.5%**	97.3%
10%	TreeShrink	198	0	100%	99.00%	100%
	Rooted pruning	200	0	100%	**100.00%**	100%

Table 1.3: Performance of TreeShrink in detecting simulated outliers. Each of the two datasets consists of 20 replicates, each has 639 HIV-1 subtype B sequences and 10 simulated outliers, for the total of 12780 subtype B HIV sequences and 200 simulated outliers.

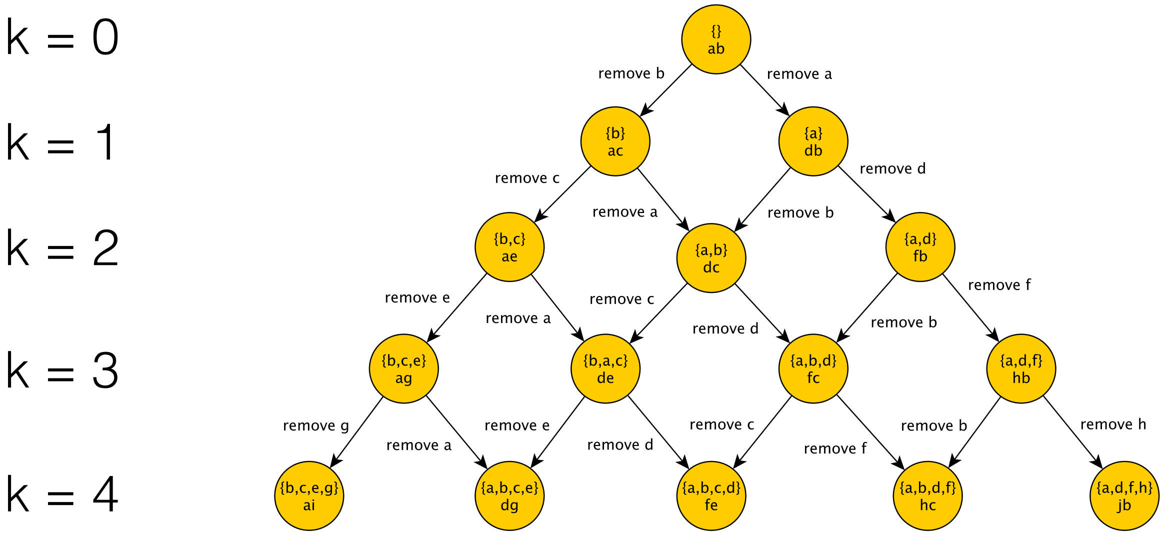
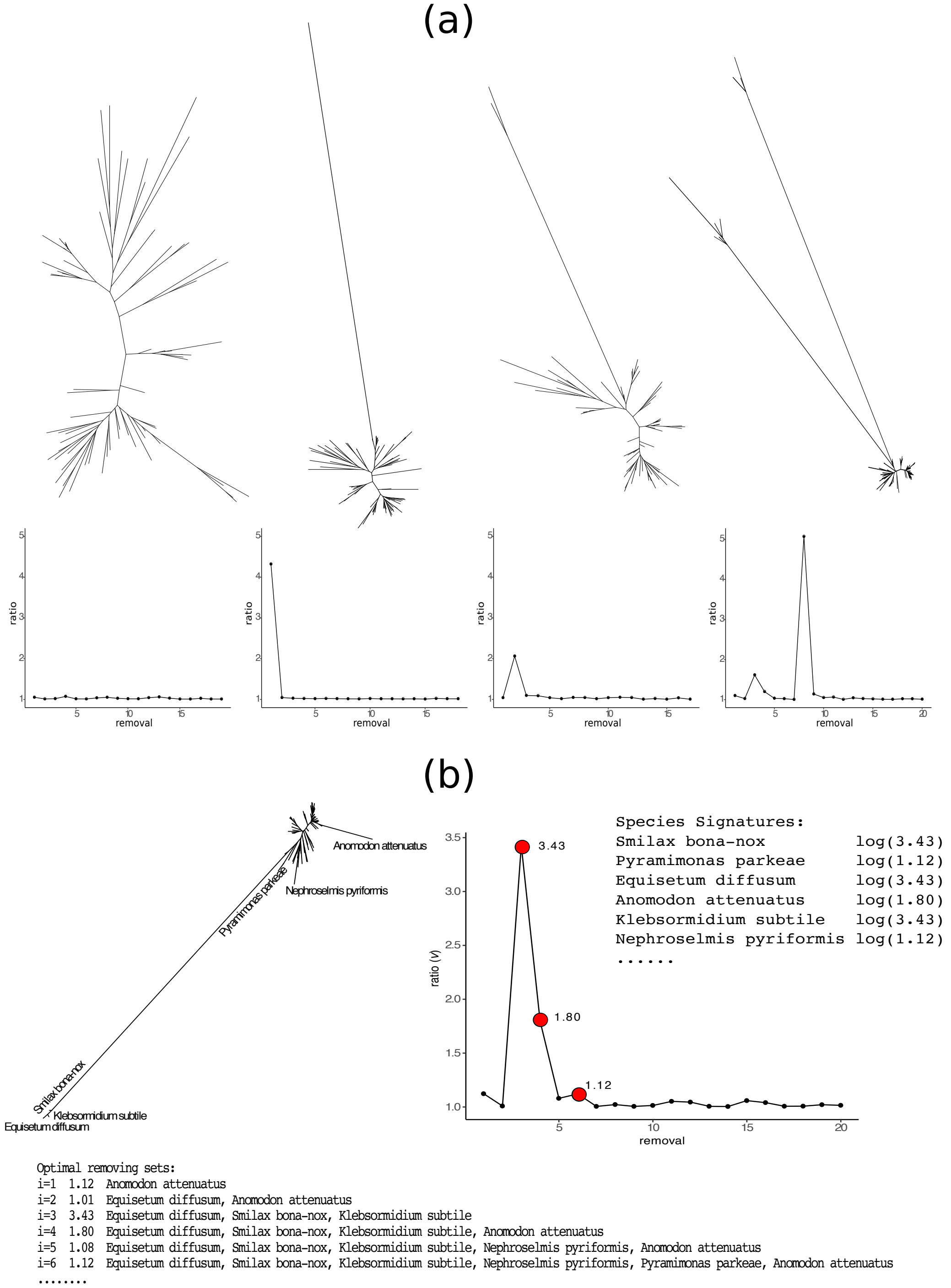


Figure 1.2: Graphical representation of the reasonable search space. The root node represents the initial tree t; each node on row k represents a restricted tree with k leaves removed. Each node is annotated by the removing set (top) and a diameter pair of the induced tree (bottom). Each edge in the graph represents a reasonable removal. The path from the root to any node corresponds to a reasonable removing chain. Each row k in the graph gives the k-removing space of t ($\mathcal{S}_k(t)$).



```
Species Signatures:
Smilax bona-nox           log(3.43)
Pyramimonas parkeae       log(1.12)
Equisetum diffusum        log(3.43)
Anomodon attenuatus       log(1.80)
Klebsormidium subtile     log(3.43)
Nephroselmis pyriformis log(1.12)
......
```

```
Optimal removing sets:
i=1  1.12  Anomodon attenuatus
i=2  1.01  Equisetum diffusum, Anomodon attenuatus
i=3  3.43  Equisetum diffusum, Smilax bona-nox, Klebsormidium subtile
i=4  1.80  Equisetum diffusum, Smilax bona-nox, Klebsormidium subtile, Anomodon attenuatus
i=5  1.08  Equisetum diffusum, Smilax bona-nox, Klebsormidium subtile, Nephroselmis pyriformis, Anomodon attenuatus
i=6  1.12  Equisetum diffusum, Smilax bona-nox, Klebsormidium subtile, Nephroselmis pyriformis, Pyramimonas parkeae, Anomodon attenuatus
........
```

Figure 1.3: (a) Patterns of ν_i as a function of i. Four unfiltered gene trees from a Plant dataset [355] are shown (*top*). For each tree, we also show ν_i for $1 \leq i \leq k = min(n/4, 5\sqrt{n})$ (*bottom*). (b) An example tree from the Plant dataset with the removing sets and species signatures. The removing sets are shown with the corresponding ν values. The max ν values associated with the species signatures are marked in red.

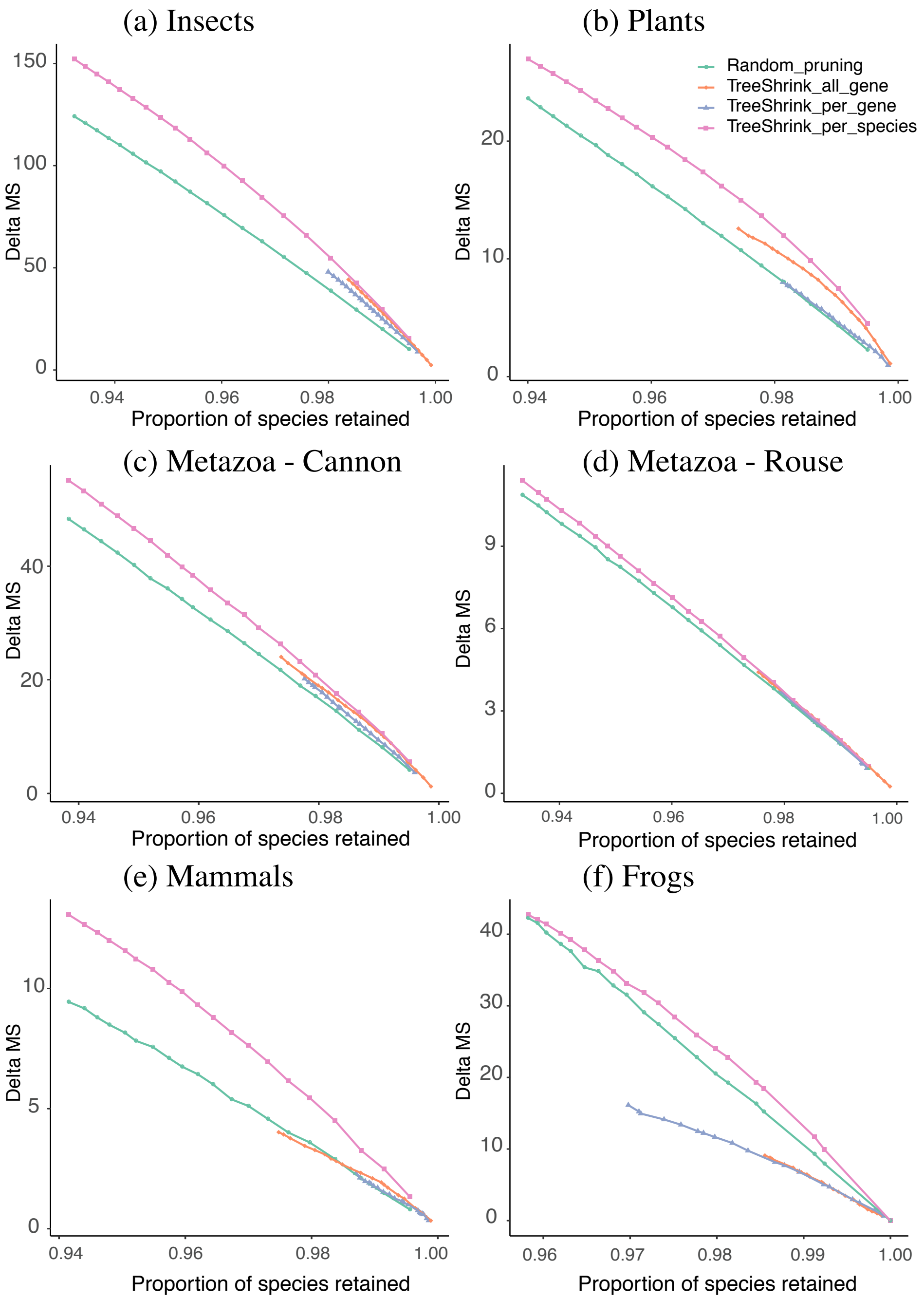


Figure 1.4: The impact of the three versions of TreeShrink on gene tree discordance on six datasets comparing to random pruning. MS distances are computed for all pairs of gene trees. The average reduction in the MS distance (*y-axis*) is shown versus the total proportion of the species retained in the gene trees after filtering (x-axis). A line is drawn between all points corresponding to different thresholds of the same method.

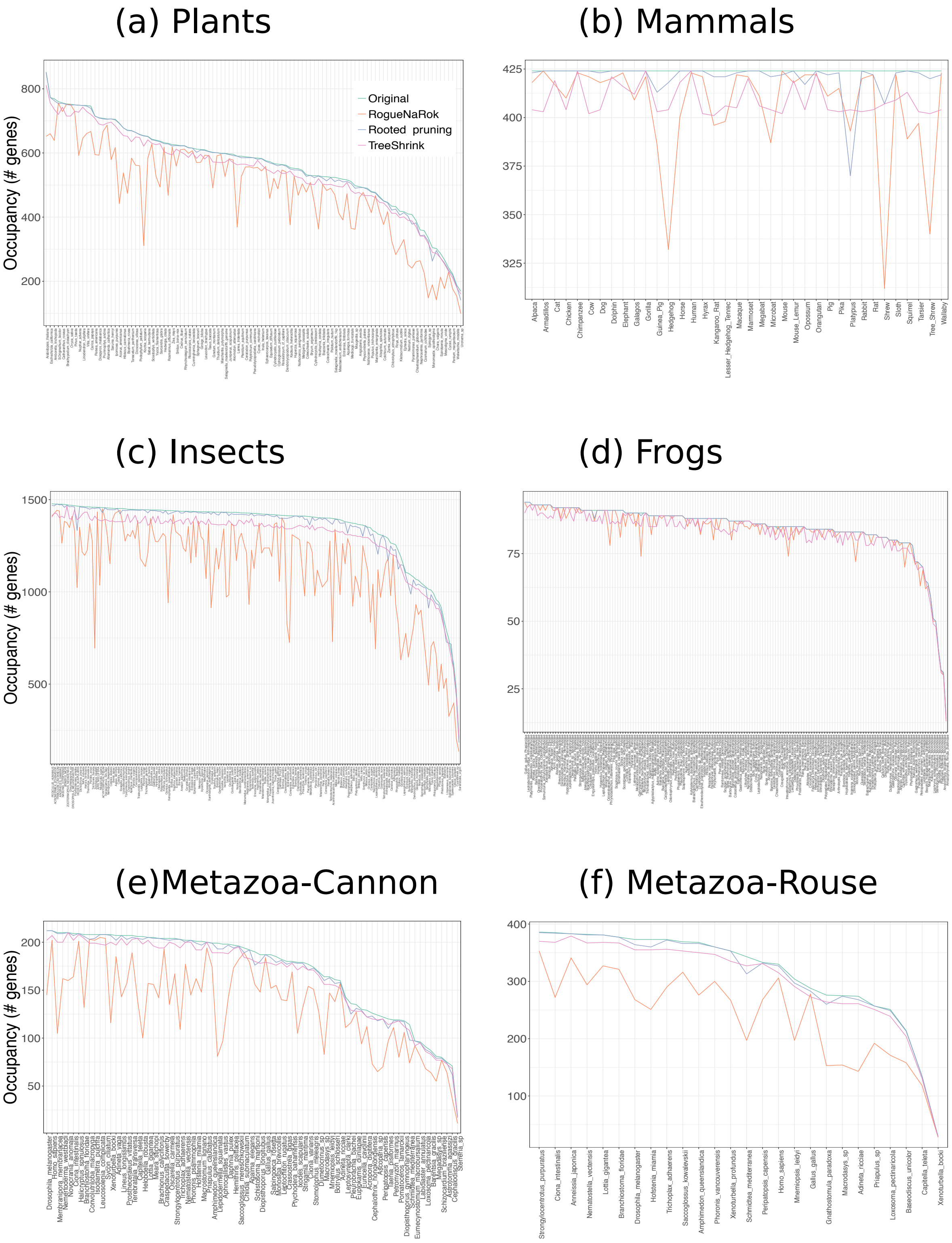


Figure 1.5: The impact of filtering on taxon occupancy for the six datasets. For each taxon (x-axis, ordered by occupancy), we show the number of genes that include it before and after filtering.

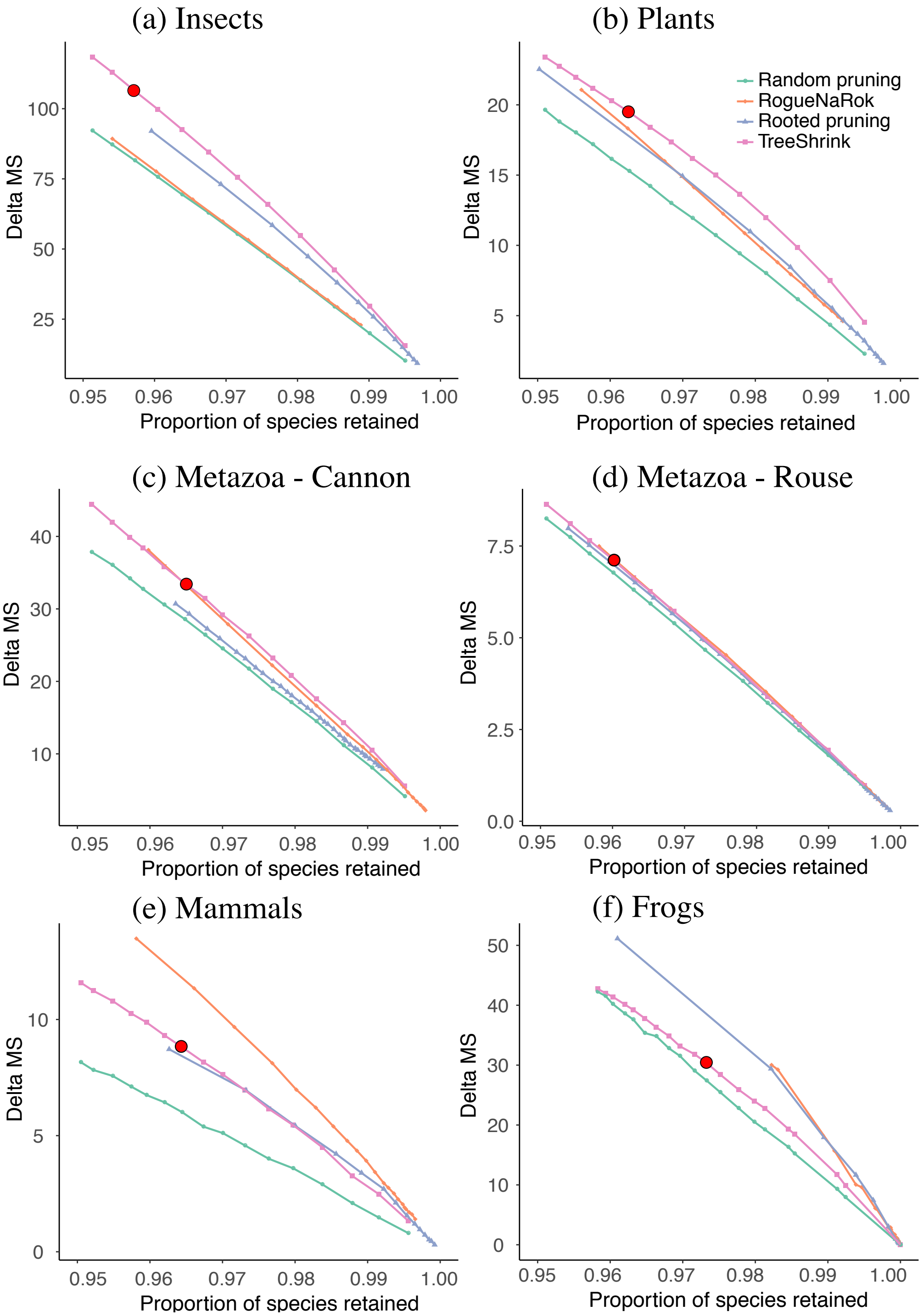


Figure 1.6: The impact of TreeShrink, RogueNaRok, and rooted pruning on gene tree discordance on six datasets comparing to random pruning. MS distances are computed for all pairs of gene trees. The average reduction in the MS distance (*y-axis*) is shown versus the total proportion of the species retained in the gene trees after filtering (x-axis). A line is drawn between all points corresponding to different thresholds of the same method. The points corresponding to the default setting of TreeShrink ($\alpha = 0.05$) are marked in red.

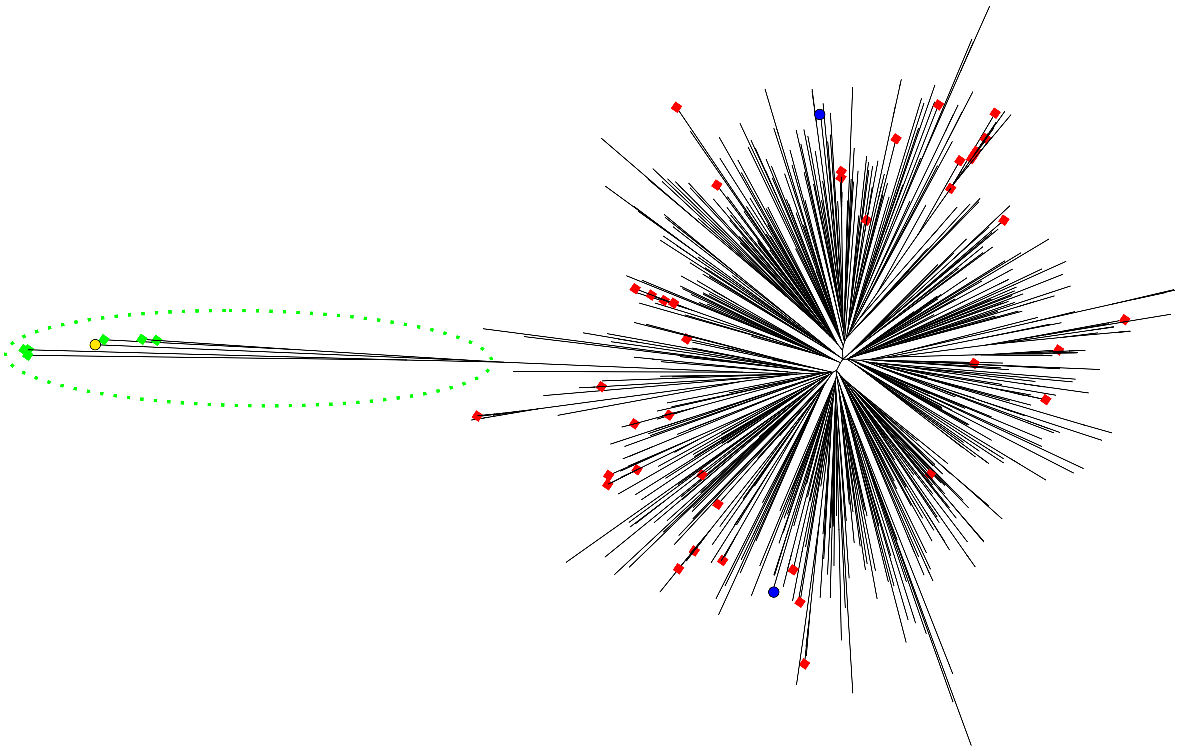


Figure 1.7: The HIV Tree. The subtype G sequence that could be detected by both TreeShrink and RogueNaRok is marked in yellow. The other non-subtype B sequences that could be detected by TreeShrink are marked in green. The subtype B species that were detected by RogueNaRok are marked in red. The two unassigned sequences are marked in blue.

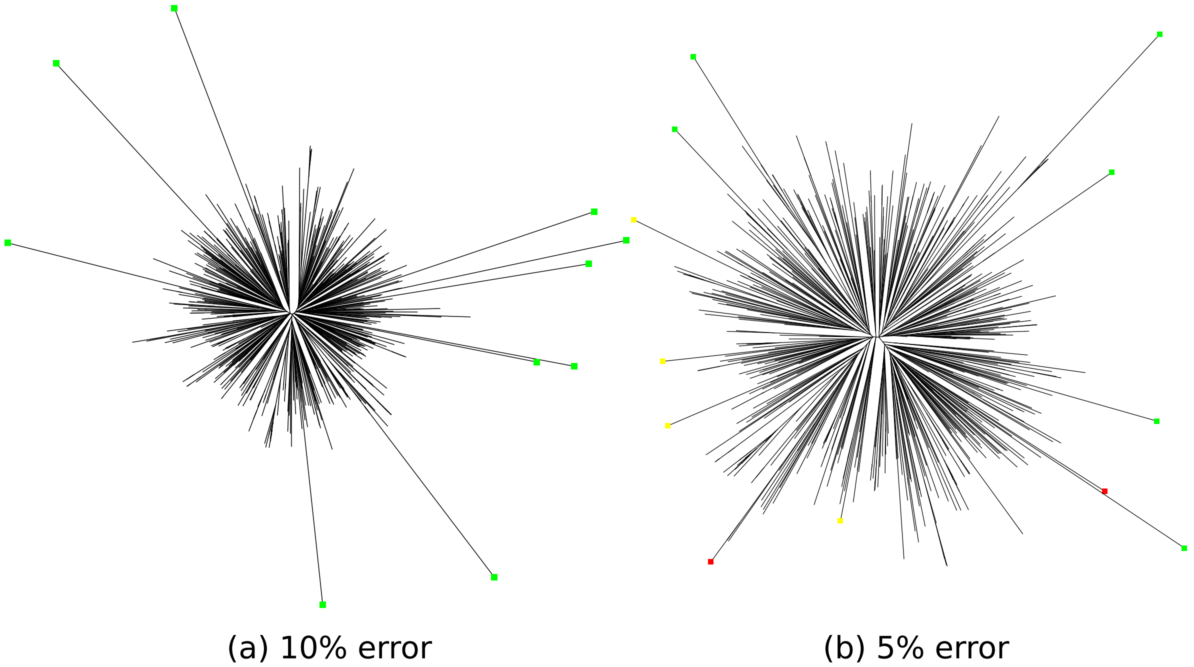


Figure 1.8: Examples of two HIV trees with 10 leaves of 10% and 5% changed in sequence. The true positives, false positives, and false negatives of TreeShrink detection (default settings) are marked in green, red, and yellow, respectively.

Chapter 2

Completing Gene Trees Without Species Trees in Sub-quadratic Time

As genome-wide reconstruction of phylogenetic trees becomes more widespread, limitations of available data are being appreciated more than ever before. One issue is that phylogenomic datasets are riddled with missing data, and gene trees, in particular, almost always lack representatives from some species otherwise available in the dataset. Since many downstream applications of gene trees require or can benefit from access to complete gene trees, it will be beneficial to algorithmically complete gene trees. Also, gene trees are often unrooted and rooting them is useful for downstream applications. While completing and rooting a gene tree with respect to a given species tree has been studied, those problems are not studied in depth when we lack such a reference species tree. We study completion of gene trees without a need for a reference species tree. We formulate an optimization problem to complete the gene trees while minimizing their quartet distance to the given set of gene trees. We extend a seminal algorithm by Brodal *et al.*, 2013 to solve this problem in quasi-linear time. In simulated studies and on a large empirical data, we show that completion of gene trees using other gene trees is relatively accurate and, unlike the case where a species tree is available, is unbiased. Our method, tripVote, is available at https://github.com/uym2/tripVote.

2.1 Introduction

Phylogenetic analyses of genome-wide data (i.e., phylogenomics) typically infer a set of gene trees, each from a different region of the genome (not necessarily a gene), and a species tree, which may be obtained from combining the gene trees. These analyses, in principle, benefit from the size of available data and can have high accuracy. However, phylogenomic datasets are also known to suffer from both partial incompleteness and undiscovered errors [180, 140, 319, 257]. The preparation of the data for phylogenomic analyses involves many steps, much of them error prone, and these steps can easily miss parts of the sequences. The issue of undetected errors is being increasingly addressed using new methods [199, 379] and simple filtering strategies [75, 298, 140]. However, by further removing data, many of these methods exacerbate the issue and sometimes have negative effects on tree inference [333, 210].

There is growing evidence that species tree inference methods are robust to presence of some missing data [141, 364, 222, 237]. The incompleteness of gene trees, however, is not just a concern for species tree inference. Gene trees are used for many other analyses, including gene family evolution, functional analyses of proteins, reconstructing ancestral gene content, and dating gene birth. Moreover, many species tree inference methods internally rely on completing gene trees, even if just approximately. For example, ASTRAL completes input gene trees with respect to each other to define a bipartitions set as its search space [219]. Thus, researchers have studied the problem of completing incomplete gene trees using the rest of the data.

The existing gene tree completion methods mostly are based on the same philosophy: that once a species tree is inferred, a gene tree can be completed with respect to that species tree to minimize their distance. What differentiates the methods is what measure of distance they use to achieve that goal. For example, [22] used a parsimony framework to minimize deep coalescence.

More recently, [53, 18] and later [4] showed how [278] (RF) distance can be minimized efficiently.

While these methods are all valuable, they do not directly provide a way to complete gene trees without a species tree. Such a completion may be desired for two reasons. First, completing gene trees with respect to the species tree may artificially reduce the amount of observed discordance. For example, if we use the species tree from the plant dataset of [241], to complete gene trees by minimizing the RF distances, the mean normalized RF distance of the gene trees to the species tree drops by 8%, meaning that the observed discordance paradoxically *reduces* as a result of gene tree completion. This level of discordance leads to an increase in estimated coalescent unit branch lengths of 0.26 on average. Thus, the added taxa are artificially less discordant with the species tree than the backbone species. Second, species trees are not always available where gene trees are. For example, part of the ASTRAL algorithm completes gene trees *before* the species tree is inferred.

We formulated gene tree completion without species trees as follows. Given a set of gene trees, complete each gene tree with respect to the other gene trees such that its overall distance to the other trees is minimized. Mathematically, the problem is similar to completion based on a species tree with the difference being that a set of reference trees (i.e., other gene trees) are available. The benefit of species-tree-free completion is that it may escape the paradoxical reduction in gene tree discordance after completion and it does not need a reference species tree. To our knowledge, very little work on this question exists. [215] introduced a method for completing gene trees by computing a quartet-based distance matrix from the gene trees and repeated use of the four-point condition. Since this heuristic method was just one step of the larger ASTRAL algorithm, it was not empirically or theoretically evaluated on its own.

Gene tree completion is also intimately connected to another problem: phylogenetic placement given a set of input gene trees. [268] introduced a method called INSTRAL for adding a new species into an species tree given a set of gene trees that already include the new species while minimizing total quartet discordance between the updated tree and the gene trees. That same mathematical problem can be used to update a gene tree using the other gene trees. When more than one taxon is missing, ordering them in some fashion and repeatedly applying the same algorithm can be used to complete the gene tree. Similarly, [267] designed a version of ASTRAL that can satisfy constraints, and the constrained version of ASTRAL can be used to complete gene trees.

In this paper, we make several contributions. First, we empirically study the species-tree-free gene tree completion problem. While past methods such as INSTRAL can be used to solve the problem, we are not aware of any study that has used them for this purpose. Second, we note that the running time of INSTRAL, which grows quadratically with the size of the backbone tree for each insertion, is sub-optimal. In a seminal work, [38] introduced an algorithm (called B13 hereafter) that allows computation of the quartet (or triplet) score between two trees in time that grows quasi-linearly with the size of the tree. Here, we extend the B13 algorithm so that it can insert new species into a tree while maximizing its total quartet score with respect to a given set of trees. Thus, we improve the asymptotic complexity of quartet-based taxon insertion (whether for gene trees or species trees). Finally, we introduce some techniques, including subsampling of quartets, that dramatically increase the accuracy of gene tree completion compared to the vanilla application of the optimization problem.

2.2 Methods

2.2.1 Notations and definitions

Let $T=(V,E)$ be a single leaf-labelled rooted tree and note all edges are directed towards a *root* node, denoted by $r(T)$. Let $e=(v,u)$ or e_v for short denote the *edge* that connects node v to u, and use $u=$ parent of v to denote that u is the *parent* of v and v a *child* of u. The set of children of an internal node $u\in V$ is denoted as $ch(u)$. We give each leaf of T a unique index in the *leafset* $L_T=\{1\dots n\}$. We use $[x]$ as shorthand for $\{0,1,\dots,x\}$. We use n_T and d_T to denote the number of leaves and the maximum degree of any nodes in tree T, respectively (omitting the subscript when clear by context). To *reroot* the tree T at an edge $e_v = (v,u) \in E$, we first divide e into two edges by adding a new vertex $r(v)$ to V and replacing e_v with edges $(u, r(v))$ and $(v, r(v))$, then we reverse the direction of all edges on the path from u to $r(T)$, and optionally, remove the old root $r(T)$ from V and make each of its children point towards its parent. The resulting graph, T_v, is a new rooted tree with the same topology as T and is called an *alternative rooting* of T on v. We use $r(v)$ to denote the root of T_v if we were to reroot T on e_v.

A *triplet* is a subtree of T induced by any three leaves in L_T (note that because T is single leaf-labeled, a triplet can be uniquely defined by a set of three leaf labels). For each triplet of T, the least common ancestor (LCA) in T of the three leaves is called the *anchor* node of that triplet. A triplet is *resolved* if the LCA of one pair of its species is not the anchor; otherwise, the triplet is *unresolved*. A *quartet* is an *unrooted* subtree of T induce by any four leaves in L_T. Note that while triplets depend on the rooting of T, quartets do not.

For two trees T_1 and T_2 whose leafsets intersect on a set S of n leaves and a given triplet $\{a,b,c\} \subset S$, we say that T_1 matches T_2 on $\{a,b,c\}$ if T_1 restricted on $\{a,b,c\}$ has identical topology to T_2 restricted on $\{a,b,c\}$. The number of *matching triplets* of T_1 and T_2 is the number of triplets that are shared among $\binom{n}{3}$ triplets on S and is denoted by $\Psi(T_1, T_2)$. For unresolved triplets, we count them as matching only if they are unresolved in both trees. Similarly, for a quartet $\{a,b,c,d\} \subset S$, and two unrooted trees T_1 and T_2, we can define $\Phi(T_1, T_2)$ as the number of quartet topologies that match between the two trees.

2.2.2 Problem formulations

We start by defining five interconnected computational problems.

Problem 1 (Maximum-matching quartet placement (MQP))**.** Given an unrooted *reference* tree R with $n+1$ leaves and an unrooted *query* tree T on all leaves of R except a leaf x, find an optimal completion T_x of T by placing x onto T to maximize $\Phi(T_x, R)$.

Problem 2 (Maximum-matching triplet rooting (MTR))**.** Given a rooted *reference* tree R and an unrooted *query* tree T, find a rooting T_v of T that maximizes $\Psi(T_v, R)$.

Problem 3 (Multi-reference MQP (m-MQP))**.** Given a collection of k *reference* trees $\mathcal{R} = \{R_1, R_2, \dots, R_k\}$ all including a leaf x (among other leaves) and a *query* tree T missing x, find a placement T_x of x on T to maximize $\sum_{i=1}^{k} \Phi(T_x, R_i)$.

Problem 4 (Multi-reference MTR (m-MTR).)**.** Given a collection of k rooted *reference* trees $\mathcal{R} = \{R_1, R_2, \dots, R_k\}$ and an unrooted *query* tree T, find a rooting T_v of T that maximizes $\sum_{i=1}^{k} \Psi(T_v, R_i)$.

The MQP problem is equivalent to the MTR problem: first root R at the taxon x, then remove x from R to obtain R', next solve MTR on R' and T to obtain the optimal rooting of T, and finally, place x on the new root of T to obtain T_x. As proved in Appendix B:

Claim 4. The tree T_x obtained by applying MTR on the query tree T using the reference tree R' and adding x as an outgroup is a solution to the MQP problem on T,x, and R.

Now consider a more general problem:

Problem 5 (Multi-reference multi-query MQP (mm-MQP))**.** Given a query tree T and a set of k *reference* trees $\mathcal{R} = \{R_1, R_2, \dots, R_k\}$ with $\bigcup_1^k L_{R_i} = L_T \cup X$, find a tree T_X on $L_T \cup X$ that is compatible with T and maximizes $\sum_{i=1}^{k} \Phi(T_X, R_i)$.

Note that the mm-MQP problem is similar to the problem solved by ASTRAL with an input constraint [267], and is NP-hard [172]. Below, we introduce algorithms to solve MTR, MQP, m-MTR, and m-MQP, and a heuristic to solve mm-MQP by sequentially applying m-MQP for each query x in X.

2.2.3 Algorithms to solve MTR, MQP, m-MTR and m-MQP

We extend the B13 algorithm to compute $\Psi(T_v, R)$ for every rooting T_v of T and select the maximum score. We first assume T and R share the same leafset of size n and then show that it is

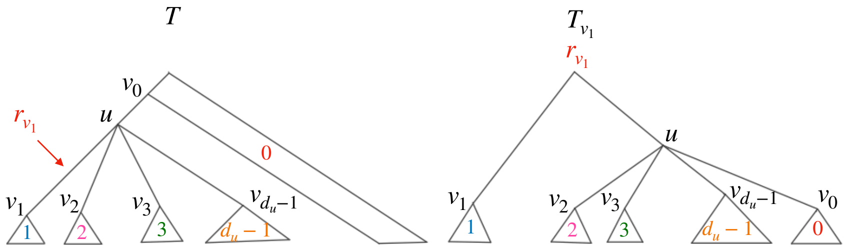


Figure 2.1: Left: The query tree T colored by node u with degree d_u. Leaves under each v_i are given the same *color* i, and the leaves outside u are colored 0. Any triplet of T anchoring at u must have two leaves taken from leaves under v_i and the other from a clade v_j different from v_i (exclude v_0 as it does not define a clade *below* u). Thus, the colors of a triplet anchoring at u must be (i,i,j) or one of its permutations, where $i \neq j, i, j \in \{1,2,\ldots,d_u - 1\}$. Right: The alternative rooting T_{v_1} of T. A new node r_{v_1} is added between u and v_1 to split the edge into two, and the edge directions are adjusted accordingly to have all nodes pointing to the new root. To count the triplets anchoring at u in T_{v_1}, we exclude v_1 instead of v_0 as in T. To count the triplets anchoring at r_{v_1}, we group all colors other than 1 into one group, and count the triplets that have colors $(i,i,1)$ or $(1,1,i)$, or a permutation of these two, where $i \neq 1$.

straightforward to extend the algorithm to trees with different leafsets.

The B13 Overview. To compute $\Psi(T_1, T_2)$, the B13 algorithm traverses T_1 top-down, and when a node u is visited, it counts the number of triplets anchoring at u in T_1 that match T_2. To do so, it colors leaves according to which side of u they belong to. To obtain the quasi-linear complexity, a *Hierarchical Decomposition Tree* (HDT) data structure representing T_2 is maintained. The HDT keeps a set of counters that allow computing the number of matching triplets for the anchor node u of T_1. HDT needs to be updated each time we move to a new node of T_1 and colors change; however, thanks to its careful design that guarantees a locally-balanced structure, updating the HDT for each leaf only takes sub-linear time.

Algorithm overview Naively using the B13 algorithm to examine each edge and choose the one with the maximum score has quasi-quadratic running time. Such a solution would be worse than that of [268], which is worst-case quadratic time. Here, we extend the B13 algorithm to solve the MTR problem in quasi-linear time (Algorithm 2). When we visit each node u in T in

Algorithm 2 Quasi-linear-time algorithm to solve the MTR problem. color(u, i) colors all the leaves below u with i. d_u is the of degree u.

```
function SOLVEMTR($T = (V,E), R$)
    HDT rooted at $\mathcal{R} \leftarrow$ BUILD HDT($R$)                    ▷ See [38]
    color(root of $T$ , 1)
    COLORANDQUERY(root of $T$)
    $\Psi$(root of $T$) $\leftarrow \sum_{u \in \text{internal nodes of } T} \tau_u^i$
    for edge $(v,u)$ in pre-order traversal of $T$ do
        $\Psi(v) = \Psi(u) - \tau_u^i - \tau_u^r + \tau_v^o + \tau_v^r$
    return $T$ rooted at $\arg\max_v \Psi(v)$
function COLORANDQUERY($u$)
    if $u$ is a leaf then
        color($u$,0) and return
    $v_1, v_2, \ldots, v_{d_u - 1} \leftarrow c(u)$
    Swap $v_1$ with the largest $v_i$ clade
    for $i = 2$ to $d_u - 1$ do
        color($v_i$, $i$)
    $\pi_0^{\mathcal{R}}, \ldots, \pi_{d_u-1}^{\mathcal{R}}, \rho^{\mathcal{R}} \leftarrow$ update HDT counters       ▷ Equ. (2.2)-(2.5)
    $\tau_u^i \leftarrow \pi_0^{\mathcal{R}}$
    $\tau_u^r \leftarrow \rho^{\mathcal{R}}$
    for $i = 1$ to $d_u - 1$ do
        $\tau_{v_i}^o \leftarrow \pi_i^{\mathcal{R}}$
    for $i = 2$ to $d_u - 1$ do
        color($v_i$, 0)
    ColorAndQuery($v_1$)
    for $i = 2$ to $d_u - 1$ do
        color($v_i$,1)
        ColorAndQuery($v_i$)
```

the topdown traversal, we compute several new counters per node (i.e., the number of triplets anchoring at u in T that match the reference tree, the number of triplets anchoring at u in *each alternative rooting* T_{v_i} for the $d-1$ children $v_1 \ldots v_{d-1}$ of u that match the reference tree, and the number of triplets anchoring at $r(u)$ in T_u that match the reference tree) that allow us to score all possible rootings. To efficiently compute these counters, we also augment the HDT with a new set of counters. Next, we first describe the node coloring scheme, then HDT and its counters, and finally our extensions.

Coloring and scoring the query tree T

Consider an arbitrary node u in T (except the root) that has degree d, $p(u) = v_0$, and $c(u) = \{v_1, \ldots, v_{d-1}\}$. The node u defines a set of d subtrees on T: the $d-1$ clades rooted at $v_1, v_2, \ldots, v_{d-1}$, and the complement subtree of the clade rooted at u. To *color* T by u, we give all leaves belonging to a same subtree of u the same color index $i \in [d-1]$. By convention, the subtree on the direction from u to the root always gets the color 0 (Fig. 2.1, left). When T is colored according to u, each triplet of T that anchors at u must have leaves with two distinct non-zero colors.

Triplet counters of the query tree To solve the MTR problem, we extend the B13 algorithm to also count the u-anchored triplets of each alternative rooting T_{v_i} of T. These triplets can be determined by the u coloring: each triplet of T_{v_i} anchored at u must have leaves colored with two distinct colors other than i (see Fig. 2.1, right). As the query tree is traversed top down in the B13 algorithm, we update it to compute and store a set of counters for each node v in T (other than the root). Let $u =$ parent of v and recall that $r(v)$ is the root of T_v; we maintain the following counters for v:

- τ_v^i: triplets *inside* v. This is the number of triplets anchored at v that match the reference tree.
- τ_v^o: triplets *outside* v. This is the number of triplets anchored at u in the alternative rooting T_v that match the reference tree.
- τ_v^r: triplets at the *rerooting* point. This is the number of triplets anchored at $r(v)$ in T_v that match the reference tree.

Note τ_v^o equals to τ_u^i in T_v (Fig. 2.1). Below, we show how to compute these counters using new HDT counters updated after each coloring of T.

Score of alternative rooting After the first top down traversal, we compute the triplet score of T (original rooting) to R by summing up τ_v^i for all nodes of T. Then, we compute the triplet score for all alternative rooting T_v of T using a second top down traversal and the following recursive formula:

$$\Psi(v) = \Psi(\text{parent of } v) - \tau^i_{\text{parent of } v} - \tau^r_{\text{parent of } v} + \tau^o_v + \tau^r_v \,. \tag{2.1}$$

Here, to move from the parent to a child, we remove matching triplets anchored at the parent or nodes outside it and add those anchored at the child or any node outside it (a triplet may be added and subtracted back).

Building and using the HDT of the reference tree

Building the HDT We use the linear-time algorithm of [38] to build the HDT data structure from the reference tree R. Each node of HDT is a combination of multiple nodes in R carefully selected in a way that ensures the HDT tree is *locally balanced*, meaning that each node with m leaves has $O(m)$ height. This local balance property enables efficient query of the number of matching triplets according to a coloring by an internal node of T. [153] refer to the nodes in the HDT as *components*, each of which is classified into one of the three types: C, G, or I (see Table B1 and Fig. B.5 in Appendix B, or refer to the original text and Fig. 2.5 of [153]). We use terms *node* and *component* interchangeably.

Updating HDT counters. To compute the number of matching triplets, each node of HDT keeps a set of counters (Table 2.1). These counters only depend on the coloring of leaves, and when a leaf changes color, the HDT counters must be updated. [38] and [153] have derived recursive formula to compute these counters for each component in the HDT given its children; thus, we can update the counters by visiting all the nodes from the leaf that has changed color to the root.

Let T be colored by node u with degree d_u and children of u, $v_1,\dots,v_{d_u-1}$ by $1,\dots,d_u-1$.

	Is the number of ...
n_i^X	i-colored leaves below X.
n_{ij}^X	pairs of leaves below X where one is colored i, the other is colored j. If X is a C type, the LCA of these two leaves must be on the external path of X and the path from the LCA to either of these two leaves does not pass through any other node on the external path. If X is a G type, the two leaves must belong to two distinct subtrees of the super root of X (Fig. B.5) in Appendix B.
$n_{i\uparrow j}^X$	pairs of leaves where one is colored i, the other is colored j, and the i-colored leaf is below the j-colored leaf in X (only defined for C type; see Fig.B.5 in Appendix B).
$n_{(ii)}^X$	pairs of leaves below X both with color i. If X is a C type, the LCA of these leaves must not belong to the external path. If X is a G type, the LCA must not be the super root of X.
$n_{(0\bullet)}^X$	pairs of leaves below X with one leaf colored 0 and the other colored different from 0 whose LCA is *not* a node on the external path if X is a C component or the super root if X is a G component.

Table 2.1: HDT Counters. Everywhere, $i, j \in [d]$. As in [153], we use the descriptors $\bullet$ and $\square$ to represent any color (unlike [153], we include 0, which is needed for rooting). Thus, $n_{i\uparrow\bullet}^X = \sum_{j\in[d]} n_{i\uparrow j}^X$; $n_{i\bullet}^X = \sum_{j\neq i} n_{ij}^X$; $n_{\bullet}^X = \sum_i n_i^X$; $n_{\bullet\square}^X = \sum_i n_{i\bullet}^X$.

Note that $d_u \leq d$ and recall that the subtree above u has color 0. In addition to counters defined by B13, we add the following two sets of counters to component X of HDT.

- ρ^X: the number of triplets of R that belong to component X and match the corresponding triplets of T_u that are anchored at r_u.
- π_j^X: the number of triplets of R that belong to component X and match the corresponding triplets of T_{v_j} that are anchored at u. If $d_u < d$, we set $\pi_j^X = 0$ for all $j > d_u - 1$.

We now show recursions for ρ^X and π_j^X and prove them in Appendix B. If X is an I or a L, we simply skip it.

If X is an IG→C, we copy over the counters of its G child.

If X is a CC→C component with children C_1 and C_2 (note that by the convention, C_1 is below C_2; see Fig. B.5 in Appendix B), then

$$\begin{aligned}\pi_j^X =& \pi_j^{C_1} + \pi_j^{C_2} + \sum_{i \neq j} \Bigg(\binom{n_i^{C_1}}{2} (n_\bullet^{C_2} - n_j^{C_2} - n_i^{C_2}) + \\ & n_i^{C_1} (n_{i\uparrow\bullet}^{C_2} - n_{i\uparrow j}^{C_2}) + (n_\bullet^{C_1} - n_j^{C_1} - n_i^{C_1}) n_{(ii)}^{C_2} + \\ & n_i^{C_1} (n_{\bullet\square}^{C_2} - n_{j\bullet}^{C_2} - n_{i\bullet}^{C_2} + n_{ij}^{C_2}) \Bigg)\end{aligned} \tag{2.2}$$

$$\begin{aligned}\rho^X = \rho^{C1} + \rho^{C2} + n_0^{C1} n_{0\uparrow\bullet}^{C2} + (n_\bullet^{C_1} - n_0^{C_1}) \sum_{i=1}^{d} n_{i\uparrow 0}^{C_2} \\ + \binom{n_0^{C1}}{2} (n_\bullet^{C_2} - n_0^{C2}) + \binom{n_\bullet^{C1} - n_0^{C_1}}{2} n_0^{C_2} \\ + (n_\bullet^{C_1} - n_0^{C_1}) n_{(00)}^{C_2} + n_0^{C_1} (n_{(\bullet\square)}^{C_2} - n_{(0\bullet)}^{C_2})\end{aligned} \tag{2.3}$$

If X is a GG→G component with children G_1 and G_2, then

$$\begin{aligned}\pi_j^X =& \pi_j^{G_1} + \pi_j^{G_2} + \sum_{i \neq j} \Big(n_i^{G_1} (n_{\bullet\square}^{G_2} - n_{j\bullet}^{G_2} - n_{i\bullet}^{G_2} + n_{ji}^{G_2}) + \\ & n_i^{G_2} (n_{\bullet\square}^{G_1} - n_{j\bullet}^{G_1} - n_{i\bullet}^{G_1} + n_{ji}^{G_1}) + \\ & n_{(ii)}^{G_1} (n_\bullet^{G_2} - n_j^{G_2} - n_i^{G_2}) + n_{(ii)}^{G_2} (n_\bullet^{G_1} - n_j^{G_1} - n_i^{G_1}) \Big)\end{aligned} \tag{2.4}$$

$$\begin{aligned}\rho^X = \rho^{G1} + \rho^{G2} + n_{00}^{G_1} (n_\bullet^{G_2} - n_0^{G_2}) + (n_{\bullet\square}^{G_1} - n_{(0\bullet)}^{G_1}) n_0^{G_2} + \\ n_{00}^{G_2} (n_\bullet^{G_1} - n_0^{G_1}) + (n_{\bullet\square}^{G_2} - n_{(0\bullet)}^{G_2}) n_0^{G_1}\end{aligned} \tag{2.5}$$

These HDT counters readily give us the $d+1$ counters associated to node u of T as defined earlier. More precisely, $\tau_u^i = \pi_0^{\mathcal{R}}$, $\tau_u^r = \rho^{\mathcal{R}}$, and $\tau_{v_j}^o = \pi_j^{\mathcal{R}}$ for each $j = [d_u]$ where $\mathcal{R}$ is the root of the HDT.

Generalizations to m-MTR and m-MQP and unequal leafsets

Note that Algorithm 2 computes and stores the score $\Psi(T_v, R)$ for *every* alternative rooting T_v of T. Thus, solving m-MTR is straightforward: we first apply Algorithm 2 to each reference tree R_i and T to compute $\Psi(T_v, R_i)$ for *all* node v of T. Then, for each node v of T, we compute $\Psi_v = \sum_{i=1}^{k} \Psi(T_v, R_i)$. Finally, we select the node v^* with maximum Ψ_{v^*} and reroot T at v^*. By Claim 4, this algorithm also solves m-MQP.

Algorithm 2 can be adopted to cases where T and R have different leafsets $L_T \cap L_R = S$ with minor modifications. Because the leaves in $L_R \setminus S$ and $L_T \setminus S$ do not contribute to the number of matching triplets, we can simply ignore them. To do so, we restrict R on S by removing from R all the leaves in $L_R \setminus S$. We mark all the leaves in T that are not in S as *inactive* and ignore the inactive leaves by not coloring them during the topdown traversal of T. The resulting algorithm is clearly correct.

2.2.4 Complexity analysis

Thank to the *smaller-half* trick of [38], at most $O(n \log n)$ leaves change color in the (recursive) topdown traversal of Algorithm 2 (i.e. the ColorAndQuery function). Therefore, the coloring module performs at most $O(n \log n)$ operations. To incorporate our extensions, three extra counters are maintained for each node in T, all of which are computed in $O(1)$ using the same topdown traversal for coloring. Thus, the asymptotic complexity of coloring does not change.

The B13 algorithm builds HDT in linear time. Because the HDT has $O(n)$ components and is locally balanced, the original HDT used in B13 can be queried in $O(\log n)$ per leaf recoloring (see [38] and [153]). Our extensions require $O(d^2)$ counters per HDT component (instead of $O(1)$ counters used in B13) increasing complexity per HDT query by a factor of $O(d^2)$. Thus, the total time complexity of Algorithm 2 is $O(d^2 n \log^2 n)$.

In a tree where d is bounded by a constant (e.g., a fully resolved binary tree), the complexity is $O(n\log^2 n)$. With k reference trees, the time complexity of both m-MTR and m-MQP is $O(kd^2 n\log^2 n)$.

2.2.5 tripVote: Completing gene trees (mm-MTR)

We develop a heuristic method Using m-MQP to complete a set of incomplete gene trees (mm-MTR). To complete a gene tree T_i, we sequentially apply the m-MQP algorithm to place each missing taxon onto T_i, using the other gene trees as references. This greedy algorithm optimizes the number of shared quartets with reference trees that include at least three of their four leaves coming from T_i. However, it does not solve the NP-Hard problem [172] of finding a complete tree with optimal quartet score over all quartets. Thus, the order of placements can change the outcome. In tripVote, we order missing taxa by their descending frequency of presence in the input gene trees, breaking ties arbitrarily. Note that tripVote only works on single-labelled gene trees.

Quartet sampling

As long appreciated, quartets with shorter terminal branches (i.e., short quartets) have better theoretical [87] and empirical [312] performance than long quartets, motivating some quartet-based methods to select short quartets [349, 230]. Inspired by these methods, we propose a stochastic approach to down-weight the votes of long quartets around the query taxon in reference trees. After rooting a gene tree at the query taxon, we sample random paths from the root to a leaf, selecting a child of a node uniformly at random at each step (Fig. B.4 in Appendix B). For each reference tree, we sample s taxa with replacement, then remove duplicates and restrict the tree to the selected set of taxa. We repeat this sampling procedure r times to generate r sampled trees for each reference tree. After sampling, we combine all the generated sample trees across all genes as the reference trees in m-MQP.

While hyper parameters s and r can be set by users, by default, we choose them such that leaves close to the root have a sufficiently high probability of being sampled at least once across the $s \times r$ draws. We first set s such that a taxon at the distance $\frac{1}{2}\log_2 n$ from the root is expected to be sampled once in each round. Since the probability of sampling a leaf at distance $\frac{1}{2}\log_2 n$ from the root in one traversal is $\frac{1}{2^{(\log_2 n)/2}} = \frac{1}{\sqrt{n}}$, setting $s = \sqrt{n}$ achieves the goal. Thus, we choose $s = \lceil \sqrt{n} \rceil$. To select the number of rounds, we set r such that a taxon with distance at most h (a predefined constant) from the root is sampled with high probability. That is, for a false negative tolerance ε, we require: $1 - (1 - \frac{1}{2^h})^{sr} > 1 - \varepsilon$ By default, we set $h = 5$ and $\varepsilon = 0.05$; thus, $s \times r = \frac{\log(0.05)}{\log(31/32)} \approx 95$ to satisfy the above inequality. Thus, by default $s = \lceil \sqrt{n} \rceil$ and $r = \frac{95}{\lceil \sqrt{n} \rceil}$.

Software package

We updated the C++ software by [289] to incorporate our algorithm to solve MTR. We built a Python wrapper, tripVote, for the C++ package and added new functions for gene tree completion using MQP, with or without the quartet sampling strategy.

2.3 Evaluation procedures

2.3.1 Datasets

We test tripVote on published simulated datasets by [219] and [201] and a real plant dataset by [241]. The simulated datasets were both created using Simphy to generate gene and species trees under the multi-species coalescent (MSC) model and heterogeneous parameters, and Indelible to simulate nucleotide sequences on gene trees according to the GTR+Γ model with varying sequence lengths and sequence evolution parameters. FastTree2 was used to estimate gene trees based on the GTR+Γ model. Original papers provide full details on the parameters used in each of these two datasets.

The **201-taxon** dataset by [219] enables us to examine the effect of incomplete lineage sorting (ILS) on gene tree completion methods. From this dataset, we use 3 model conditions with 200 taxa, and use the first 20 out of the 50 replicates of the original dataset (to save computational time). In each replicate, we use the first 500 (out of 1000) estimated gene trees that are fully-resolved. The three model conditions have medium, high, and very high levels of ILS, resulting in 21%, 33%, and 69% RF distance between true gene trees and the species tree. They also have high levels of gene tree estimation error (15%, 22%, and 34% RF between estimated and true gene trees).

The **31-taxon** dataset by [201] is used to examine the effect of clock deviation on gene tree completion methods. Here, we use the 3 model conditions with the root to crown ratio equal to 1.0, but varying levels of deviation from the clock (low, medium, high). We only use the first 20 out of the 100 replicates of the original dataset because our experiments are computationally intensive. The average coefficient of variation of root-to-tip distances of low, medium, and high deviations are 0.04, 0.30, and 0.69, respectively. These replicates have moderately high level of ILS (with 24% mean RF distance between true gene trees and the species tree). The amount of gene tree estimation error increases with deviations from the clock (RF error are 30%, 41%, and 52%).

The real OneKP biological dataset of 1178 plants by [241] has 384 gene trees, all of which miss some of the species. The original study provide an ASTRAL species tree inferred from 384 gene trees, inferred using RAxML, each with at least $384/2 = 192$ species.

2.3.2 Experiments

We compare tripVote with two alternatives tree completion algorithms: *ASTRAL-completion*, the method used in ASTRAL and described in [215], and *OCTAL*, the gene tree completion method that minimizes RF distance of each gene tree to the species tree. ASTRAL-completion is run using the ASTRAL software, and OCTAL is run using the TRACTION-RF software [53]. In addition, to guide visualization and interpretation, we add a lower-bound control by randomly completing the gene trees (repeated 100 times and averaged).

In simulated datasets, we randomly remove m leaves from each estimated gene tree to create incomplete gene trees; $m \in \{0,1,2,20,50,100\}$ for the 201-taxon dataset and $m \in \{0,1,2,3,8,15\}$ for the 31-taxon dataset. We use tripVote, OCTAL, and ASTRAL-completion to complete each set. For the 201-taxon dataset with $m = 1$, we compare the accuracy of tripVote with and without the sampling.

We use two different error metric: the normalized *RF* distance and the *induced RF distance*, as described below. To separate the gene tree estimation error from the error by completion methods, we define the *induced* (normalized) RF distance, as follow: given two trees T_1, T_2 and a reference tree R, the induced RF distance of T_2 on T_1 with respect to R is $\frac{RF(T_2,R)-RF(T_1,R)}{RF(T_1,R)}$ where RF denotes the normalized RF distance of two trees after restricting them to their shared leafset. Positive (negative) induced RF distances show that T_1 (T_2) is closer to the reference tree. On the simulated datasets, we use the estimated gene tree by FastTree as T_1, the tree completed by a completion method (e.g. ASTRAL-completion, tripVote, etc.) as T_2, and the true gene tree as R.

In addition, we test the ability of tripVote to improve species tree estimation. On the 201-taxon dataset, we compare five versions of ASTRAL for inferring species trees from incomplete gene trees. (1) The default ASTRAL uses ASTRAL-completion to construct the search space and original trees to score. (2) We use tripVote in place of ASTRAL-completion but continue to score trees using incomplete trees. (3) We use OCTAL in place of ASTRAL-completion. Since running OCTAL needs a species tree, we use the ASTRAL species tree inferred in (1) as input to OCTAL. Thus, in this setting, ASTRAL is run twice. (4) We use the gene trees completed by ASTRAL-completion as input to ASTRAL, making them used both for search space creation and scoring. (5) Similarly, we use tripVote completed trees as input. We measure the error of these ASTRAL trees by computing their RF distances to the true species tree.

We also test tripVote and ASTRAL-completion on the ability to root an unrooted gene tree with respect to other rooted gene trees. On the two simulated datasets, we remove the outgroup from a set of $n-k$ gene trees (arbitrarily selected) and use the k remaining trees to infer the outgroup placement. We vary k in $\{1,10,50,100,250,500\}$. To measure error, we compute the optimal rooting that minimizes the triplet distance to the true tree and report the *delta triplet error*, defined as the difference between the triplet distances of a rooted tree and the optimal tree to the true tree. In addition to ASTRAL-completion, we also compare tripVote to other rooting methods, including the outgroup rooting (root at the original placement of the outgroup before removing it), mid-point rooting, and MinVar rooting [201] and add the random rooting as a control.

For the OneKP dataset, we set up two versions: one where the *original* gene trees are used directly and one with *extra missing data* where we prune out an extra *p%* of the taxa from each gene tree (for $p \in \{5,10,15,20\}$). With original data, where the completed gene trees are unknown, we measure the induced RF distance of the completed gene tree (T_2) on the original (incomplete) one (T_1) with respect to the species tree (R). For the extra missing data, after running the methods to complete the gene trees, we reduce the completed gene trees to the same leafset as the original gene trees and compute their normalized RF distances.

2.4 Results

2.4.1 Simulated datasets

Gene tree completion

Across all model conditions with $m = 1$, the sub-sampling strategy dramatically lowers the error compared to full quartet sampling (Fig. B.6 in Appendix B). Both versions of tripVote have higher error when the ILS level increases. The averaged error of tripVote with and without sampling are 0.51 versus 0.84, 0.77 versus 1.75, and 3.77 versus 5.42 for medium, high, and very high ILS, respectively. Thus, the error is less than half for the high ILS level and is reduced everywhere. Looking beyond the average error and examining the full distribution shows that while in the majority of cases error is at most one branch with sampling, the same is not true when sampling is not performed. Both versions suffer from a tail of placements with very high error (e.g., five edges or more), a condition that unsurprisingly is observed mostly for the highest level of ILS. However, the tail of large errors is clearly reduced after sampling. Because restricting the calculations to shorter quartets has a clear positive impact on the results, we use sampling by default in tripVote and use it in the remaining experiments.

Comparing tripVote and ASTRAL-completion, across all conditions, tripVote always has lower error, and the difference between the two methods is more pronounced when the number of missing taxa increases (Fig. 2.2). The relative improvements of tripVote compared to ASTRAL-completion are quite large. On the 201-taxon dataset at 50% missing data (i.e. $m = 100$), the induced RF error of tripVote is 32%, 34%, and 6% lower than that of ASTRAL-completion in medium, high, and very high ILS levels, respectively (Fig. 2.2b). Similarly, tripVote dominates ASTRAL-completion on the 31-taxon dataset across all conditions of clock deviation, albeit with smaller differences compared to 201-taxon dataset. For example, with $m = 15$, the induced RF error of tripVote is 11%, 2%, and 4% lower in low, medium, and high clock deviations, respectively (Fig. 2.2d).

The comparison between tripVote and OCTAL depends on the dataset and the level of missing data. On the 31-taxon dataset, tripVote has better accuracy, and the improvements are most pronounced with higher clock deviations and medium level of missing data (e.g., eight taxa). On the other hand, OCTAL is more accurate in most conditions of the 201-taxon dataset and especially when the amount of missing data exceeds 50 taxa. Improvements of OCTAL over tripVote are non-existent or negligible for the highest levels of ILS and are increased for lower levels.

All methods are affected by the level of missing data, ILS (Fig. 2.2a,b), and clock deviations (Figure 2.2c,d). Even before completion, gene trees have higher levels of errors when the ILS is higher or when the deviations from the clock are more pronounced. Completion always increases error compared to estimated gene trees, and increases in the error are higher when there are more missing data. However, this increase in error is more pronounced for the highest level of ILS than lower levels. Thus, for very high ILS, not only gene tree estimation is difficult, completion is also difficult. In particular, RF distances after completion can reach 0.7 for the highest ILS case. In contrast, average levels of RF remain below 0.33 after completion for medium or high ILS. Thus, gene tree-based completion using tripVote fails to be accurate at the highest levels of ILS. In contrast to ILS levels, we did not detect a reduction in effectiveness of tripVote as deviations from the clock increase. In fact, induced RF errors go down with increasing levels of clock deviations (Fig. 2.2d). Note that with high deviations, the error is already very high before completion and there is relatively little room left for increased error.

Effects on species tree accuracy

We ran ASTRAL to infer the species trees from incomplete gene trees under five different settings (described earlier) and compared their normalized RF errors (Fig. 2.3). All ways of running ASTRAL showed *some* level of sensitivity to missing data, especially for high ILS and with more than 50 missing taxa per gene (i.e., $\approx 25\%$ of the leaves). In contrast, the condition with the lowest level of ILS is remarkably robust to even extreme levels of missing data ($\approx 50\%$ of the leaves).

At all levels of ILS, the accuracy is always higher when the completed gene trees are only used to construct search space than when they are also used for scoring species trees. Overall, the best accuracy is obtained when tripVote is used only for building the search space. In this setting, tripVote slightly improves upon the default ASTRAL-completion method when ILS is very high and there is moderate amount of missing data (i.e. up to 50 taxa). Thus, tripVote can be used in place of ASTRAL-completion inside ASTRAL to improve its accuracy. Moreover, tripVote has far better accuracy than ASTRAL-complete when the completed gene trees are used both for search space and scoring. Comparing to the original setting of ASTRAL (which uses ASTRAL-completion for search space only), the ASTRAL tree inferred using OCTAL either has the same accuracy (when ILS is medium) or worse. Note that the OCTAL setting uses the default ASTRAL species as input. Therefore, our results do not show any benefit in using OCTAL for improving the search space of ASTRAL.

Gene tree rooting

The accuracy of tripVote for rooting is mixed. The absolute error of tripVote rooting clearly increases with the level of ILS (Fig. 2.4 Top, but not with deviations from the clock (Fig. 2.4 Bottom). The accuracy of tripVote (and also ASTRAL-completion) rapidly increases as the number of voting trees increases to 100, but there is relatively little improvement after that. Overall, the accuracy of tripVote improves as a result of adding the sampling strategy; however, the improvements are more subtle than those observed for the placement problem.

In all model conditions, tripVote is more accurate than ASTRAL-completion, but its accuracy comparing to other methods depends on the model condition. With medium ILS, tripVote is the best method and even outperforms outgroup rooting (Fig 2.4a). With high ILS, tripVote is similar in accuracy to MinVar. However, when ILS is very high, tripVote is not a good choice (Fig. 2.4a). Overall, if an outgroup is available, it is clearly a better choice than tripVote when ILS levels are high or very high. When clock deviation is low, branch-length-based rooting methods are very accurate and better than outgroups and tripVote. (Fig 2.4b). In medium clock, the error of MinVar and MidPoint go up but still slightly dominate tripVote, and outgroup rooting is the most accurate. When the clock deviation is high, MinVar and MidPoint have higher error, and tripVote is the best method given enough number of voting trees (Fig. 2.4b).

Running time

We note that tripVote, if run without the sampling strategy to complete the species tree, solves a similar problem to INSTRAL. Using the dataset from the original study by [268], we compare the running time of INSTRAL and the two versions of tripVote with and without sampling (Fig. B.2 in Appendix B). Here, the species tree (not a gene tree) is being completed, and the two methods are guaranteed to return the same solution; the only difference is the running time. The running times of the two methods are comparable when they both use complete input gene trees as input. The theoretical running time of INSTRAL depends on the number of *unique* nodes across all gene trees (e.g., tripartitions for a binary tree), and thus, very similar gene trees do not increase its running time dramatically. However, in practice, gene trees often miss at least *some* leaves, forcing most nodes to be distinct. Thus, we also tested a case where gene trees missed only 1% of the leaves. Under these conditions, tripVote is much faster. For example, with 10000 species, INSTRAL takes on average 71 minutes while tripVote takes only 14 minutes.

Consistent with the theory, the asymptotic running time of INSTRAL grows faster than linearly (close to $n^{1.4}$) without missing data and close to quadratically with missing data. In contrast, tripVote running time without sampling increases close to linearly with or without missing data. With sampling, because we set the sampling size to a sublinear function of n, the running time of tripVote further reduces and is sublinear (close to $n^{0.9}$).

2.4.2 Real datasets

On the real dataset, we show the incongruence of the completed gene trees (original data) with the species tree (Fig. 2.5a), the error of the completed gene trees at different levels of extra missing data (Fig. 2.5b), and the estimated branch length of the species tree (for original and extra missing data at 20%, Fig. 2.5c). Consistent with the results of simulated data, here we also see that tripVote is more accurate than ASTRAL-complete, but is not as accurate as OCTAL, especially with higher levels of missing data. Thus, in terms of topological accuracy of gene trees alone, using the species tree to complete the gene trees gives the best results.

The completed OCTAL trees, however, are biased. Ideally, the completed trees should be no more or less distant to the species tree than the original incomplete trees, and the induced RF distance should be distributed around 0. Both the random and the OCTAL methods substantially change the distance to the species tree, especially when the number of missing taxa increases. The random completion sharply increases the induced RF distance with a high variance. While the induced RF distance of the OCTAL method has very low variance at all levels of missing data, the value decreases below 0 when the missing data increases. This reduction shows that the OCTAL method produces completed gene trees that have *lower* discordance with the species tree than incomplete gene trees, and indicates that the resulting completed trees may be overfit to the species tree. ASTRAL-completion and tripVote have relatively little effect on induced RF distance and keep it around 0 even at the highest levels of missing data.

The two methods have the opposite tendencies: ASTRAL-completion tends to slightly increase the distance to species tree (mean induced RF: 0.035) while tripVote tends to slightly decrease the distance (mean induced RF: -0.015). Also, ASTRAL-completion has a higher variance compared to tripVote (0.006 versus 0.002).

As a result of these biases, when OCTAL-completed gene trees are used to estimate the species tree, the OCTAL trees cause an overestimation in the species tree branch lengths compared to using the original gene trees (Fig. 2.5c). Such a problem is far less severe when tripVote or ASTRAL-completion is used. Both original data and the extra 20% missing data show a consistent pattern. As expected, the branches of the species tree estimated using random completed gene trees are underestimated compared to the original branch lengths obtained from incomplete gene trees.

2.5 Discussions

We introduced a quasi-linear algorithm for adding a new taxon to a tree to maximize its total matching quartets to a given set of reference trees that already include the taxon. We built a method called tripVote around this algorithm by using a sampling strategy to further improve accuracy and a simple greedy algorithm to allow adding multiple taxa. Overall, results indicate that species-tree-free completion of gene trees does add to the error of the trees, compared to what could be achieved if sequences were available. This much should not be surprising. Gene tree based completion was also not always more accurate than species tree aware completion. However, results indicate that gene-tree-based completion is able to maintain the overall levels of gene tree discordance with the species tree. Thus, unlike species tree aware completion, the method does not seem biased toward increasing or decreasing the gene tree discordance. Two main factors limiting the accuracy of gene tree completion seem to be the true levels of gene tree discordance (e.g., ILS) and the amount of gene tree error (controlled in our experiments using deviations from the clock).

Comparing the species tree aware method, OCTAL, and tripVote, we saw mixed results. While tripVote has better accuracy with higher clock deviations and moderate levels of missing data, OCTAL is more accurate in other settings, and the gap increases with the number of missing taxa. While part of the differences may be due to the inherent advantage of using a species tree, the more subtle issue of optimality needs to be also considered. While OCTAL is an exact algorithms that minimizes the RF distance of each gene tree to the species tree, tripVote is a greedy heuristic when there are more than one missing taxon. Its heuristic nature may explain why tripVote's accuracy degrades with the level of missing data more quickly than OCTAL, as its error after each m-MQP application can add up. Note that since OCTAL requires a species tree to operate, it has two limitations: it makes the completed gene trees biased towards the species tree used and it is not useful for the species tree estimation problem (even in the 2-iteration setting where we tested it). In contrast, tripVote works directly on a set of gene trees and maintains a more faithful distribution of the gene trees discordance after completion. Therefore, tripVote is more suitable than OCTAL in use cases that need to maintain the discordance level or have to avoid the use of a reference tree.

While we tested tripVote for gene tree completion, the MQP and *m*-MQP algorithms can be used in other contexts such as species tree completion. In that usage, tripVote (without sampling) would be identical to INSTRAL in terms of the resulting placement (both solve the same problem exactly) but will have a better worst-case running time complexity. This better running time also opens the door for developing methods that can infer an entire species tree by repeated placements (i.e., using a greedy algorithm to solve an NP-Hard problem). While a simple greedy search may not outperform methods such as ASTRAL [378], repeated applications of the greedy search may provide a better running time versus optimality trade-off. We leave the exploration of such directions to future work. gn,

The ability of tripVote to root trees was mixed and depended on the dataset. Given the difficulty of knowing the model condition on real data, we do not necessarily advocate using tripVote for rooting, unless when researchers know the levels of ILS are not high and *some* deviations from the clock are expected. Otherwise, using methods such as MinVar seems preferable. Future works can improve the rooting accuracy by combining tripVote and branch-length-based rooting. One direction could be incorporating the MinVar score of each branch in addition to the tripVote score, but that approach requires a way to combine the two scores. Taking the idea further, machine learning techniques could perhaps be used to combine the scores from multiple methods to find the best rooting overall by training for parameters of a function that combines these scores as features.

The tripVote method can also be improved in several ways. First, since tripVote is a greedy algorithm, the ordering of the taxa to be inserted may affect its accuracy. Future works can explore different strategies to order the queries or run multiple times and summarize results across multiple orderings. Second, the current setting gives the same weight to the vote of every reference tree, regardless of its distance to the backbone tree. As the topology of different gene trees can vary substantially, a weighting scheme that discounts the votes of distant gene trees to the backbone tree should be explored. Lastly, while tripVote computes all the individual votes of every reference tree, it only uses their sum to select the placement branch. Another research direction is to explore other strategies to summarize the votes, such as using the median, or a non-linear transformation of each triplet score before summing. Alternatively, one can also take a machine learning approach to use the set of votes from the reference trees as features to learn and predict the best placement branch, in a framework such as that of [152].

2.6 Acknowledgements

Chapter 2, in full, is a reprint of the material as it appears in “Completing Gene Trees Without Species Trees in Sub-quadratic Time”. Mai, Uyen, and Siavash Mirarab. Bioinformatics (2022). The dissertation author was the primary investigator and first author of this paper.

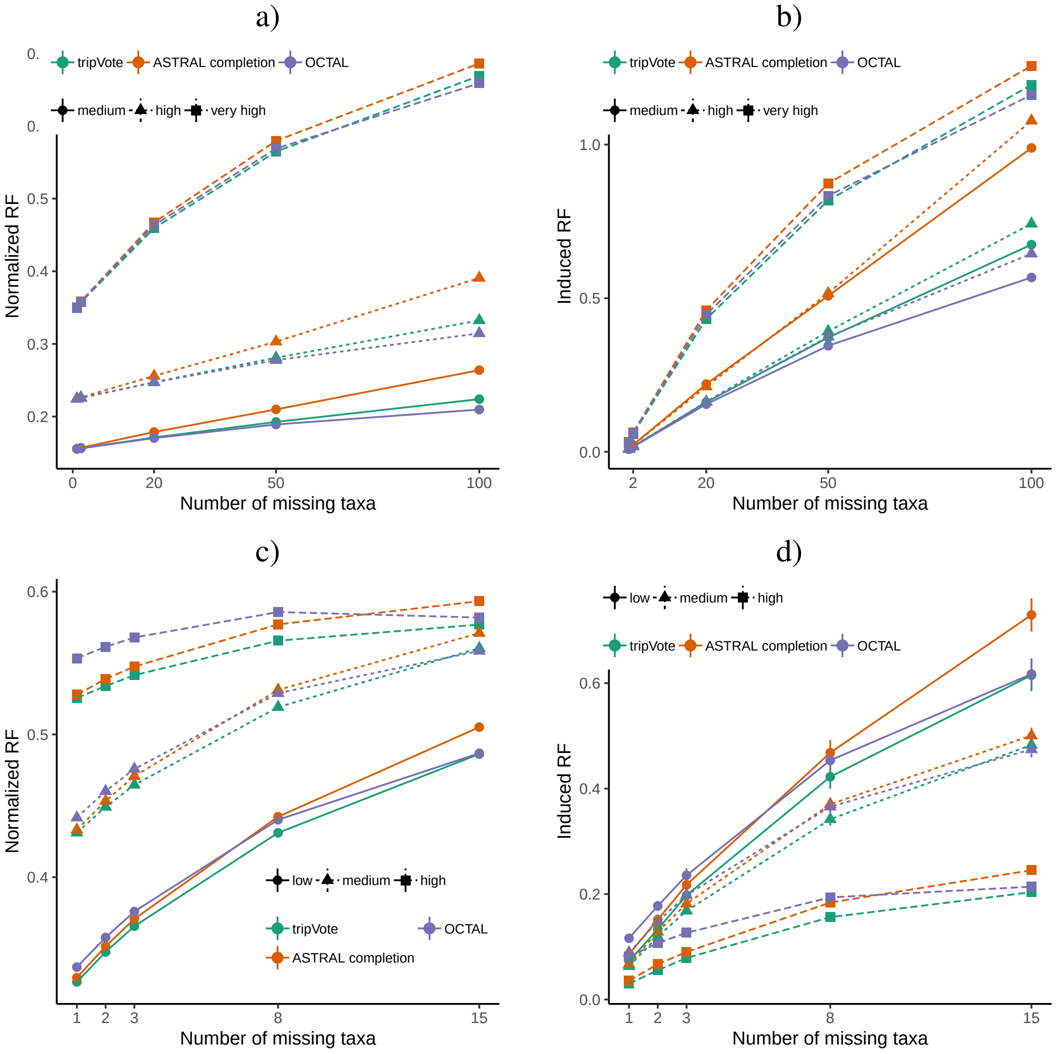


Figure 2.2: (a,c) Normalized RF error of tripVote, OCTAL, and ASTRAL-completion on the 201-taxon dataset with different levels of ILS, (a), and the 31-taxon dataset with different levels of clock deviations, (c); $m = 0$ shows the average RF error of the complete gene trees estimated by FastTree. (b,d) Induced RF error of tripVote, OCTAL, and ASTRAL-completion on the 201-taxon dataset, (b), and the 31-taxon dataset, (d). Also see Figures B.8 and B.9 for a different view that includes the random completion as control.

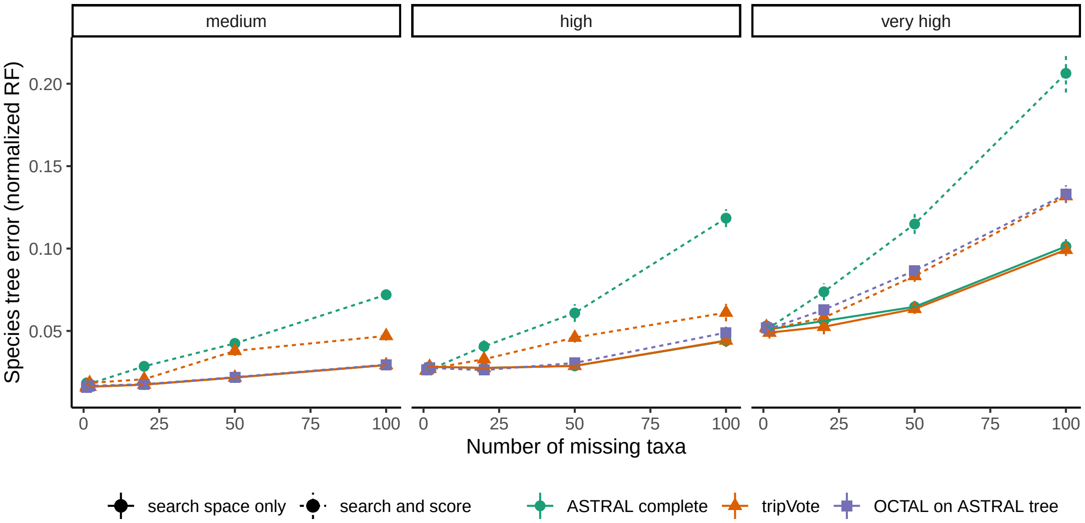


Figure 2.3: Topological error of the ASTRAL species tree estimated using different set of gene trees (the 201-taxon dataset). The three panels show different levels of ILS. In “search space only”, the completed gene trees (by ASTRAL-complete, tripVote, or OCTAL) were only used to guide ASTRAL’s search space, whereas in “search and score”, the completed gene trees were used as the actual input to ASTRAL. To obtain the results for OCTAL, two rounds of ASTRAL was run: in the first round the search space was produced by ASTRAL-complete and the incomplete trees were used as input; in the second round, ASTRAL was run using the OCTAL-completed gene trees, both for search space and as input.

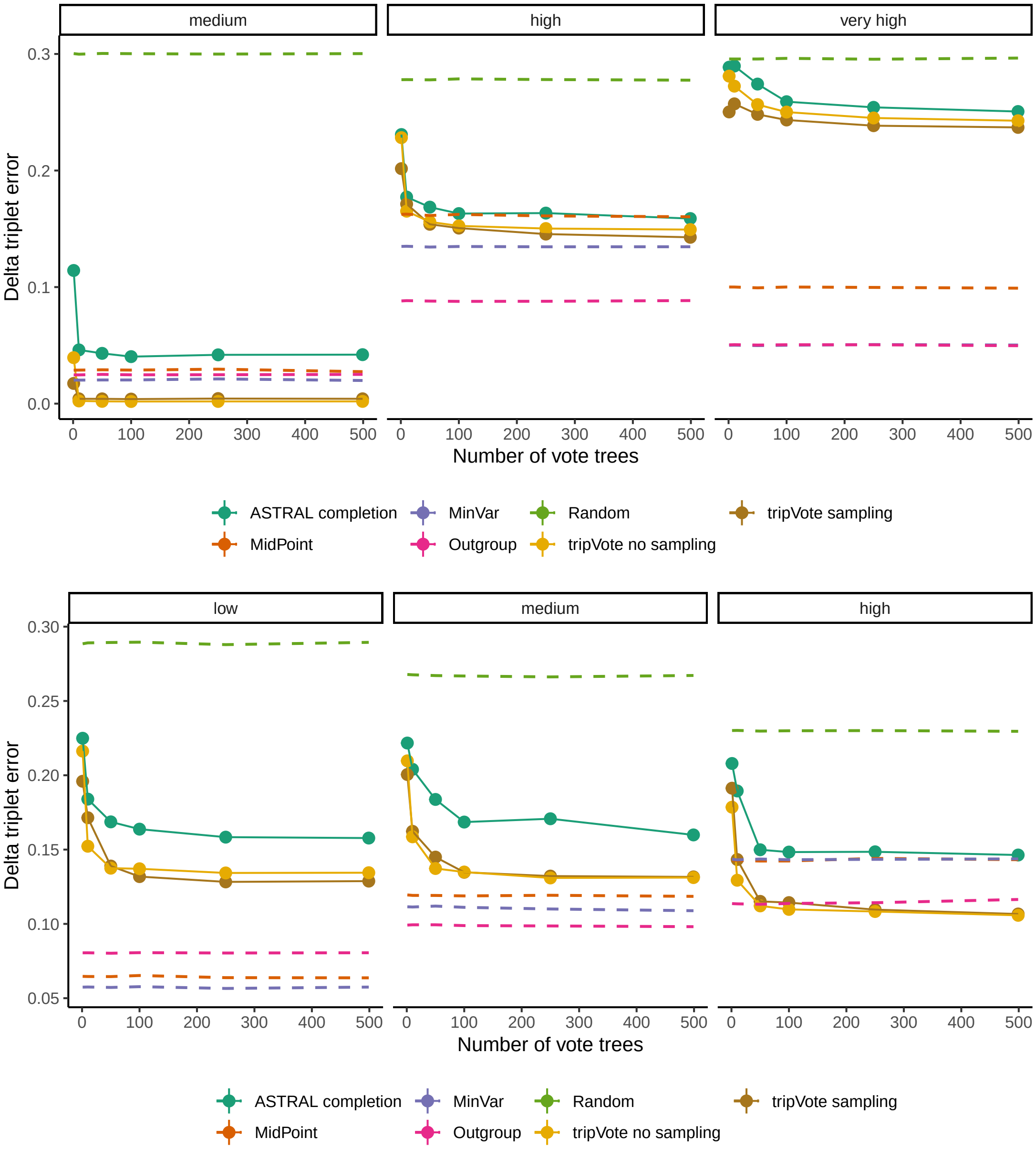


Figure 2.4: Accuracy of rooting based on different methods for Top: The 201-taxon dataset and Bottom: The 31-taxon dataset. The outgroup is removed from m randomly selected trees and inserted back using either ASTRAL-completion or tripVote, then each of these trees is rerooted at the reinserted outgroup. The x-axis shows the number of voting trees for ASTRAL-completion and tripVote (i.e. $n-m$) and the y-axis shows the delta triplet error (i.e. the triplet error to the true rooted tree subtracting the triplet error of the optimal rooting that has minimum triplet error to the true tree). We added alternative rooting methods (Outgroup, MinVar, MidPoint, and Random) that do not use other gene trees. Outgroup rooting was done on the complete estimated trees with outgroup included. MidPoint and MinVar were run *after* the outgroup was removed. The Random rooting was repeated 50 times and the average error is reported. See also Fig. B.3 in Appendix B where the error is measured by the raw triplet error.

a) Species tree discordance (original data)b) Gene tree error (extra missing)c) Species tree branch lengths.

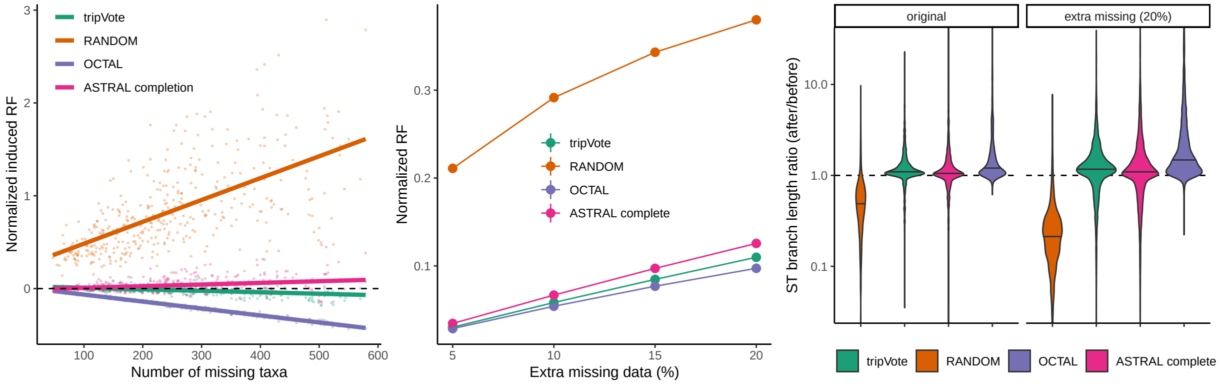


Figure 2.5: The OneKP results for (a and left panel of c) completing incomplete gene trees, and (b and right panel of c) completing gene trees with extra (introduced) missing data. (a) Induced RF distance to the species tree of different completion methods on the original incomplete gene trees versus the number of missing taxa. (b) The ratio of the species tree branch lengths after versus before completion by different methods; the y-axis is shown in logarithmic scale. See Fig. B.7 in Appendix B for normalized RF and another view of the branch length estimation.

Chapter 3

Minimum variance rooting of phylogenetic trees and implications for species tree reconstruction

Phylogenetic trees inferred using commonly-used models of sequence evolution are unrooted, but the root position matters both for interpretation and downstream applications. This issue has been long recognized; however, whether the potential for discordance between the species tree and gene trees impacts methods of rooting a phylogenetic tree has not been extensively studied. In this paper, we introduce a new method of rooting a tree based on its branch length distribution; our method, which minimizes the variance of root to tip distances, is inspired by the traditional midpoint rerooting and is justified when deviations from the strict molecular clock are random. Like midpoint rerooting, the method can be implemented in a linear time algorithm. In extensive simulations that consider discordance between gene trees and the species tree, we show that the new method is more accurate than midpoint rerooting, but its relative accuracy compared to using outgroups to root gene trees depends on the size of the dataset and levels of deviations from the strict clock. We show high levels of error for all methods of rooting estimated gene trees due to factors that include effects of gene tree discordance, deviations from the clock, and gene tree estimation error. Our simulations, however, did not reveal significant differences between two equivalent methods for species tree estimation that use rooted and unrooted input, namely, STAR and NJst. Nevertheless, our results point to limitations of existing scalable rooting methods.

3.1 Introduction

Commonly-used models of sequence evolution, such as GTR [336], are time reversible and can therefore be used to reconstruct *unrooted* phylogenetic trees. The correct placement of the root is often of intrinsic interest as evident by long debates on the correct rooting of the universal tree-of-life [173, 41, 100, 322, 115, 274], and other major groups (e.g., [206, 335, 322]). Moreover, the knowledge of the root is often needed for downstream applications of phylogenetic trees, such as ancestral state reconstruction [197], comparative genomics [92], taxonomic profiling of metagenomic samples [232, 207], and dating.

Several approaches have been proposed for this long recognized issue [251]. The current prevailing practice is to simply use outgroups [197]. An outgroup is a species known *apriori* to be outside the group of interest (referred to as the ingroup). Outgroup selection is an art that requires balancing two opposite goals; the outgroup needs to be divergent enough from the ingroup to make its outgroup status unambiguous, but at the same time not so distant that strong long branch attraction [256, 131, 24] negatively impacts the resolution of the ingroup, the placement of the outgroup, or both [334, 116, 137, 187, 281]. Nevertheless, several studies have found outgroups to be competitive with more complex methods [142, 35] that use evidence from molecular data for rooting.

At one end of the spectrum, rooting an unrooted tree is trivial when the rooted tree is ultrametric (i.e., all leaves are equidistant to the root). Only one rooting of an unrooted tree can create an ultrametric tree, and that rooting can be obtained by midpoint (MP) rooting; i.e., root the tree at the middle point of the longest path between any two leaves of the tree. A phylogenetic tree with branch lengths measured in the expected number of mutations will be expected to be close to ultrametric if mutations follow a strict molecular clock (i.e., rates of mutation are constant). When a strict molecular clock is not followed in the data, one can still use the midpoint rooting, hoping that divergences from a strict clock are small and that midpoint rooting can still be a good proxy for the correct root [129]. At the other end of the spectrum, non-reversible models of sequence evolution, such as the General Markov Model[20, 9, 150], or those that incorporate nonstationarity [27, 32], can be used to infer a rooted tree from the data; however, these methods have not yet enjoyed broad application because of statistical issues related to model complexity and lack of scalability to large datasets (but see [357] for recent advances).

Despite the long history of thinking about tree rooting, we believe the question should be revisited in the phylogenomic era. The potential for discordance among gene trees and incongruence with the species tree due to factors such as incomplete lineage sorting (ILS) is now well-understood [196, 68, 83] and many empirical analyses strive to account for it [149, 355, 314, 263, 13] (but see [106, 311, 318, 84] for the ongoing debate on this issue).

Rooting phylogenies needs fresh thinking in the phylogenomic area for several reasons. Firstly, an outgroup is a species believed to be outside the ingroups in the true species tree; however, depending on how the outgroup is chosen, its true position may or may not be outside the ingroups in every single gene tree. As an example, according to the multi-species coalescent model [244], an outgroup separated from the ingroups by a branch of length 2 in coalescent units [68] (corresponding to 8 millions years assuming a diploid effective population size of 200,000 and a generation time of 10 years) is expected to be mixed with the ingroups in 9% of genes only because of ILS effects and optimistically assuming that all basal branches of the ingroups are so long that only two lineages have to coalesce in the branch below the root. Thus, even if outgroups are reliable methods of rooting a species tree, they may fail to root every gene tree accurately. A second reason to revisit rooting is related to the practice of species tree estimation. The most scalable pipeline for estimating a species tree first estimates a set of gene trees and then uses a "summary method" to combine the estimated gene trees to reconstruct a species tree. Some summary methods (e.g., MP-EST [192], STAR/STEAC [193], and GLASS [224]) rely on rooted input gene trees, while more recent methods (e.g, ASTRAL [218, 219] and NJst [191, 343] – also known as USTAR/NJ [8]) can combine unrooted gene trees. Even though the question has never been directly addressed before, the accuracy of methods based on unrooted trees tends to be superior to rooted trees on simulated and empirical data [212, 219, 218, 191, 311]. It remains to be tested if these trends relate to incorrect rooting of gene trees, as suggested by some studies [311]. Finally, reconciliation between gene trees and a species tree may provide a way to root them. Gene duplication history, the number of deep coalescences, and distributions of unrooted gene trees have all been used to root gene trees, species trees, or both [163, 33, 7, 249, 374]. However, in this manuscript, we will focus on rooting gene trees individually and not collectively or with reference to a known species tree.

Beyond phylogenomics, the ever-expanding size of phylogenetic trees is another factor that should be considered in discussions of rooting. Trees with thousands of leaves are routinely inferred and used currently, and trees with many hundreds of thousands of leaves are also in use [322, 132, 266, 209]. We should ask whether rooting such large trees with existing methods is computationally feasible, and if so, whether they are accurate.

In this paper, we address the problem of rooting large phylogenomic datasets. We introduce a new rooting method that minimizes the variance of the root to tip distances. We implement our new method, called min-var (MV) rooting, in an algorithm that scales linearly with the tree size, just like the MP rooting (note that the term minimum variance used here does not relate to statistical minimum variance estimators). We compare MV and MP with outgroup (OG) rooting under a wide range of conditions where gene trees and the species tree can be discordant, with a range of dataset sizes, with several ways of choosing an outgroup, and with various levels of divergence from a strict clock. We then go on to compare several species tree reconstruction methods, including those that use inferred unrooted trees, or trees rooted using the three rooting approaches. Our rooting tool is publicly available at `https://uym2.github.io/MinVar-Rooting/`.

3.2 Materials and Methods

3.2.1 Min-Var (MV) rooting

Notations and definitions

Let an unrooted tree be represented as a connected acyclic undirected graph $G = (V, E)$, and let each edge $e = (u, v) \in E$ be weighted by a length w_e. To root G at an edge $e = (u, v) \in E$ and a position $x \leq w_e$ from u, we first divide e to two edges by a vertex p and replace e with edges (p, u) and (p, v) with lengths x and $w_e - x$, respectively. Then, we convert G to a directed graph by pointing all its edges away from p. The resulting graph is a rooted tree, T, and is a *rooting* of G.

We use the following notations for a rooted tree T. Each node u in T, except the root $r(T)$, has a parent, $p(u)$, and the child set of a node u is denoted by $c(u)$. A node u is either *internal* and has two or more children or is a *leaf* and has no children. The set of leaves is denoted by $L = \{1 \dots n\}$. For any node u, we denote the length of the edge $(p(u),u)$ by e_u. For each point p on this edge (including u), we let $Cld(p)$ denote the set of leaves descending from node u and $|p|$ is used for the size of $Cld(p)$. For two points p and p', potentially on different edges, we let $d(p, p')$ denote the total length of the *undirected* path from p to p', and use $d_i(p) = d(i, p)$ as a shorthand for $i \in L$.

We set $mean(p) = 1/n \sum_{i \in L} d_i(p)$, $var(p) = 1/n \sum_{i \in L} (d_i(p) - mean(p))^2$, $SI(p) = \sum_{i \in Cld(p)} d_i(p)$, and $ST(p) = \sum_{i \in L} d_i(p)$.

We call a p_0 a *local MV* of tree T if and only if for any point p and $x = d(p_0, p)$,

$$\lim_{x \to 0} \frac{var(p) - var(p_0)}{x} = 0 \tag{3.1}$$

and the second derivative of $var(p_0)$ is non-negative (i.e., $var(p) > var(p_0)$).

The *global MV* of a tree is a point p_0 that has the minimum $var(p_0)$ among all positions on all branches of the tree. Unless otherwise specified, we use the terminology *MV* to refer to the *global MV*.

A point p is said to be a *balance point* of T if the average of tip distances to p are equal on any two sides of p in the unrooted version of T; that is, p is a balance point if

$$\frac{1}{|u|} \sum_{i \in Cld(u)} d_i(p) = \frac{1}{n - |u|} \sum_{i \notin Cld(u)} d_i(p) \tag{3.2}$$

for all ways of choosing u such that p is on the edge $(p(u), u)$ (including both ends).

Problem statement

MP rooting can be framed as an optimization problem that seeks the rooting position that minimizes the maximum distance from any leaf to the root. Our proposed approach, MV rooting, is based on a similar idea, but minimizes the variance instead of the maximum.

The MV problem is: *Given* an unrooted tree G, *find* a rooting T^* of G such that

$$T^* = \arg\min_T var(r(T)) \ . \tag{3.3}$$

Thus, we seek the root that minimizes the variance of root to tip distances.

Motivation for MV rooting

We start with the following propositions (proofs are shown in Appendix C)

Proposition 3. *A point p on tree T is a local MV if and only if it is a balance point.*

Based on Proposition 3, we refer to local MV and balance point interchangeably.

Proposition 4. *Any tree has at least one local MV.*

Proposition 5. *The global MV of any tree is one of its local MVs.*

When the strict molecular clock is followed, the true rooted phylogenetic tree is ultrametric with zero root-to-tip distance variance. For ultrametric trees, only the true rooting position is a balance point, and therefore, the tree has a unique local MV at the correct root, which is also its global MV (Proposition 5). Since local MVs are also balance points, they provide a natural choice for rooting when there are randomly distributed deviations from the molecular clock. Among several local MVs, the global MV also minimizes the total variance, and arguably is the best choice. We now describe a simplified model under which we can prove that MV is in expectation the correct root.

Random deviations model: Consider a model where a rooted tree T is generated from an ultrametric tree T_0 by multiplying the length of each edge (u,v) by a random variable α_v drawn from any distribution with support $[1-\varepsilon, 1+\varepsilon]$ and expected value 1. Let h be the height of T_0 and r be the position of the true root on T, which it inherited from T_0. We have the following two propositions (proofs are shown in Appendix C).

Proposition 6. *Let p denote the global MV of T. If*

$$\varepsilon \leq \min_{w \in c(r)} \left(\frac{e_w}{\frac{n}{n-|w|}h + e_w} \right)$$

then there exists a child w of r such that $p \in e = (r,w)$

Following Proposition 6, the global MV is guaranteed to be on one of the adjacent edges of r if ε is sufficiently small. Note that the restriction on ε is a sufficient but not a necessary condition. Regardless of the value of ε, we can also show the following.

Proposition 7. *When the global MV is on one of the adjacent edges of r, let a random variable X indicate the distance of the global MV to the root; then,* $E(X) = 0$.

Corollary 3. *Under our random deviations model where deviations from the strict molecular clock are independent and bounded, the MV rooting will find the correct branch, and in expectation, will also have zero distance on that branch to the correct rooting position.*

Although the random deviations model considerably simplifies real biological processes, it is useful in motivating the MV rooting approach in general.

The MV rooting algorithm

The algorithm is based on the following proposition (proof is shown in Appendix C)

Proposition 8. *Let p be a point on an edge* (u,v) *of tree T with distance* $d(p,u) = x$*. If we let p vary along edge* (u,v) *and consider* $var(p)$ *as a function of variable x with parameters u and v, then:*

$$var(p) = var(x;u,v) = (1-\beta^2)x^2 + \left(\alpha - \frac{2ST(u)\beta}{n}\right)x + var(u) \tag{3.4}$$

in which

$$\alpha = \frac{2ST(u) - 4(SI(v) + |v|e_v)}{n} \quad \text{and} \quad \beta = 1 - \frac{2|v|}{n} \tag{3.5}$$

To find the MV root, we first arbitrarily root the unrooted tree at r_T to get a rooted tree T. We then use Algorithm 3 to traverse T three times to search for local MVs. At the end, we select the local minimum with the lowest variance value as the global MV.

Traversal 1 and 2 (Preprocessing): In the first top down traversal, we trivially compute the distance to root (i.e., $d(u,r_T)$) for all nodes of the tree, and then simply compute the variance of root-to-tip distances. Next, in a post-order traversal, for each node u, we compute the size of its clade (i.e., $|u|$) and the sum of distances to the tips in its clade (i.e., $SI(u)$), both of which are simple to compute.

Algorithm 3 Linear time MinVar rooting algorithm

```
Function MinVarRoot($T$)
For node $u$ in pre-order($T$)                                   # Top-down traversal
    Compute $d(u,r(T)) = d(p(u),r_T) + e_u$ for $i \in L \backslash \{r(T)\}$
$minvar \leftarrow \sigma^2(\{d_i(r(T)) | i \in L\})$                         # $\sigma^2$ is variance
For node $u$ in post-order($T$)                                  # Bottom-up traversal
    Store $|u| = 1$ if $u \in L$, else $|u| = \sum_{v \in c(u)} |v|$
    Store $SI(u) = 0$ if $u \in L$, else $SI(u) = \sum_{v \in c(u)} (SI(v) + e_v|v|)$
    $globalMV \leftarrow r(T)$
For node $v$ and $u = p(v)$ in pre-order($T$)                      # Top-down traversal
    Compute and store $ST(v)$ using Eq. 3.6
    Compute and store $var(v)$ using Eq. 3.4
    Compute $x^*$ using Eq. 3.7 and call the corresponding point $p^*$
    Compute $var(p^*)$ using Eq. 3.4
    if $minvar > var(p^*)$
        $minvar \leftarrow var(p^*)$
        $globalMV \leftarrow p^*$
    reroot $T$ at $p^*$
```

Traversal 3: The final top-down traversal finds the local MV along each edge (u, v) if it exists, and records the local MV with the minimum root-to-tip variance as the global MV. We set $ST(r_T) = SI(r_T)$ and for other nodes we compute and store:

$$ST(v) = ST(p(v)) + (n - 2|v|)e_v. \tag{3.6}$$

According to Proposition 8, for any point p along the edge (u, v) with $x = d(u, p)$, we can compute $var(p)$ (the variance of root-to-tip distance if we root at p) using Eq 3.4. Let $a = (1 - \beta^2)$, $b = (\alpha - \frac{2ST(u)\beta}{n})$, and c = $var(u)$; Eq. 3.4 is a standard quadratic function $ax^2 + bx + c$ with $a > 0$ (because $|\beta| < 1$) and with the restriction $x \in [0, e_v]$. Thus, $var(p)$ is minimized on a point p^* with distance x^* from u where:

$$x^* = \begin{cases} \frac{-b}{2a}, & \text{if } \frac{-b}{2a} \in [0, e_v] \\ \arg\min_{x \in \{0, e_v\}} (var(x; u, v)), & \text{otherwise} \end{cases} \tag{3.7}$$

and p^* is a local MV of T only if $x^* = \frac{-b}{2a}$. Since we compute this for all edges, at the end, we have all local MVs and their corresponding root-to-tip variance; we simply select the local MV point that has the lowest variance and reroot T on that point. Derivation of Eq. 3.6 and the proof for Proposition 8 are shown in Appendix C.

Theorem 5. *Algorithm 3 is guaranteed to find the global MV.*

Proof. It is clear that Equation 3.7 minimizes 3.4 given the constraint $x \in [0, e_v]$ (recall that the second derivative $a > 0$) and thus finds local MV points. According to Proposition 8, Equation 3.4 gives the correct variance of root-to-tip distances for any point on the tree. By the definition of global MV and Propositions 4 and 5, the global MV p is always the local MV with the minimum $var(p)$. Because Algorithm 3 checks all edges for all local MVs and compute root-to-tip variance at all of those points, it guarantees to find the correct global MV. □

Proposition 9. *The running time of Algorithm 3 scales linearly with the number of leaves in the tree.*

Proof. Algorithm 3 visits each edge in T exactly three times, each of which involves only constant time operations. After the rooting position is found, rerooting the tree also takes no more than linear time assuming that the tree is represented with the usual pointer structure. Thus, the overall time complexity of Algorithm 3 is O(n). □

Algorithm 4 Linear time midpoint rooting algorithm.

```
Function MidpointRoot($T$)
For node $u$ in post-order($T$)                                   # Bottom-up traversal
    $MI(u) \leftarrow \max(\{MI(v) + e_v | v \in c(u)\})$
$MO(r(T)) \leftarrow 0$
For node $v$ in pre-order($T$)                                    # Top-down traversal
    $MO \leftarrow \max(\{MO(p(v))\} \cup \{MI(s) + e_s | s \in c(p(v)) - \{v\}\})$
    $x^* \leftarrow (MI(v) - MO + e_v)/2$
    if $x^* \geq 0$ and $x^* \leq e_v$
        reroot $T$ at $(u, v)$ with distance $x^*$ from $u$ and return
    $MO(v) \leftarrow MO + e_v$
```

Similar to MV, MP rooting can be done in linear time using two tree traversals (Algorithm 4). Interestingly, at least one phylogenetic package in common use, Dendropy, seems to have opted not to implement this simple algorithm, and instead uses an approach that scales quadratically with *n* (our attempt to use ape [245] failed). We re-implemented MP using the Dendropy package to solve this shortcoming.

3.2.2 Experimental design

Simulated datasets

We study four simulated datasets, including two that were previously published. One of the published datasets, RNASim [217], includes only one gene tree and is used here only to evaluate the scalability of rooting methods. The other datasets all use SimPhy [202] to generate gene and species trees under the multi-species coalescent (MSC) model [244] and heterogeneous parameters. We then used Indelible [98] to simulate nucleotide sequence evolution on gene trees according to the GTR+Γ model with varying sequence length and different sequence evolution parameters (Appendix C). Then, FastTree2 [262] was used to estimate gene trees based on the GTR+Γ model.

The three main datasets with species trees and gene trees are:

- D1 – 30-taxon heterogeneous dataset: Here, the number of ingroup species was fixed to 30. We simulated 100 replicates, each with a different species tree and 500 gene trees. This dataset is used for extensive analysis of all methods.

- D2 – Large heterogeneous dataset: This dataset includes two subsets, one with 2000 and another with 5000 taxa, and is used for testing performance on large datasets. For both datasets we created 20 replicates with different species trees and 50 gene trees.

- D3 – ASTRAL-II dataset: We reused a previously published dataset [219] to investigate performance for intermediate number of species (i.e, 10, 50, 100, 200, 500, and 1000).

The new datasets, D1 and D2 are simulated using a similar approach. For each number of species in both D1 and D2, we simulated 10 different model conditions where we changed parameters that control divergence from the strict clock and the distance of the outgroup to the ingroups. Seven out of ten model conditions included an outgroup. The outgroup is added as a sister to the ingroups on the species tree. The length of the branch above ingroups (connecting them with the root) is decided by multiplying a fixed number by the height of the ingroup species tree; we refer to that fixed number as the root to crown ratio (R/C). For example, an R/C of 0.5 indicates that the branch connecting the root of the ingroups to the root of the tree is half the height of the ingroup tree. The choice of the R/C ratio directly impacts how often the species tree outgroup is also a gene tree outgroup (Fig. 3.1C).

Beyond the R/C ratio, model conditions are also distinguished by two parameters of SimPhy that control deviations from the clock: (i) gene by lineage specific rate heterogeneity, which is a multiplier drawn from a gamma distribution for each branch of each gene tree, and (ii) species specific branch rate heterogeneity rate, which is also a multiplier drawn from a gamma distribution per species and is used to scale all gene tree branches for that species universally. The gamma distributions are mean-preserving, and therefore are specified with one shape parameter. We draw the value of that shape parameter from a log normal distribution with the scale hyperparameter $\sigma = 1$ and a varying location hyperparameter, which controls the level of deviation from the strict clock. We refer to the log normal location (which is the log of the mean of the distribution minus 0.5) as the clock deviation parameter; the higher values correspond to gamma distributions more closely centered around one, and thus, less deviation from the clock, while lower values correspond to more deviation (Fig. 3.1D).

In six model conditions, the clock deviation parameter is fixed to a moderate value of 1.5, and the R/C ratio is varied between 0 (no outgroup), 0.25, 0.5, 1, 2, and 4. In the remaining model conditions, the R/C ratio is fixed to either 0 or 1 and the clock deviation parameter is changed between 0.15, 1.5, and 5 to get high, moderate, and low levels of deviations, respectively. Note that the two model conditions with heterogeneity hyperparameter 1.5 are common with six conditions that varied the R/C ratio; thus, in total we have ten model conditions for each number of species.

Other parameters of the SimPhy simulation procedure are sampled from distributions as described in Appendix C. In D2, the expected species tree height in is set to 14.7 million generations, which is much higher than the 3 million used for the 30-taxon dataset. We chose different heights for small and large datasets because having 30 surviving species in a span of 3 million generations is reasonable, but having many thousands of extant species in such a short evolutionary time is unlikely. Thus, for the D2 dataset, we increased the height to obtain more realistic conditions.

The portion of quartet trees induced by gene trees that are found in the species tree can be used as a measure of ILS [297], where values close to 1/3 indicate extremely high levels of ILS and values close to 1 indicate no ILS. Our datasets varied between these two extremes (Fig. 3.1A). The gene tree estimation error, measured by RF distance between true gene trees and estimated gene trees, was similarly heterogeneous and was also substantially impacted by deviations from the clock (Fig. 3.1B); with low and medium deviations, median gene tree error was respectively 25% and 32%, while for high deviations, the error increased to 49%.

A major point of the current paper is that an outgroup species is not always an outgroup in gene trees, even in the *true* gene trees. When the R/C ratio is low, many of the true gene trees do not have the outgroup species in the outgroup position (Fig. 3.1C). Interestingly, with the 30-taxon dataset, only at the extremely high R/C= 4 the outgroup is outside the ingroups in close to all gene trees of all replicate datasets. At the other extreme, with R/C= 0.25, in more than 50% of replicate runs, more than 50% of our 500 true gene trees did not have the outgroup species in the outgroup position. The larger datasets, which had higher numbers of generation and higher levels of ILS (Fig. 3.1A) had fewer cases of outgroup mixing with ingroups in true gene trees.

Evaluation metrics

To estimate the accuracy of a rooted gene tree, we measure the proportion of all $\binom{n}{3}$ triplets in the reference (i.e., true) tree that are also found in the estimated tree. This measure is a function of both the accuracy of the unrooted topology of the estimated tree and the accuracy of the rooting. To separate the rooting error from the tree error, for the small 30-taxon dataset where it is feasible, we examine all possible root placements and find the "ideal" rooting that results in the lowest possible triplet distance (the ideal triplet distance is zero if and only if the unrooted tree is correct). We then define "delta triplet distance" as the difference between the triplet distance of the estimated tree with the rooting of interest (OG/MV/MP) and the triplet distance of the estimated tree with the ideal rooting. For the small trees with 30 leaves, we also afford to compute the rooted SPR distance using SPRDist [363]; however, for larger trees, SPR could not be computed. Finally, for the true unrooted gene trees that are rooted using an algorithm, we also report normalized branch distance, defined as the number of branches between the correct root and the estimated root, normalized by the maximum number of branches from any leaf to the root.

Beyond triplet distance, we use the normalized RF distance to measure the accuracy of unrooted trees, and we use percentage of quartets in the gene trees also present in the true gene trees (as computed by ASTRAL [218]) as a measure of ILS. For species trees, we also report the Matching Split measure [28]. We also report running time, measured on Intel EM64T Xeon nodes with 64GB memory.

Implementations

We implemented both MP and MV (`https://uym2.github.io/MinVar-Rooting/`) using the Dendropy package for phylogenetic manipulations [327]. As expected, the running time of the algorithm increases linearly with the number of leaves (Fig. 3.2); an RNASim [217] tree with 200,000 leaves could be rooted in just under a minute. In contrast, Dendropy seems to use a quadratic implementation of MP rooting (Fig. 3.2A).

3.3 Results

We will examine the following research questions using simulated and empirical data:

- *RQ1:* Does our novel MV rooting improve the root placement accuracy compared to MP and OG rooting for datasets with varying numbers of species?
- *RQ2:* How are MP, MV, and OG impacted by (i) gene tree estimation error, (ii), divergence from the clock, and (iii) outgroup distance to ingroups (R/C)?
- *RQ3:* What is the impact of rooting error on the species tree estimation, and is STAR less accurate than its unrooted counterpart, NJst?

3.3.1 Simulation results

RQ1: MV for varying numbers of leaves

On the D1 (30-taxon) dataset with estimated gene trees, MV matched or improved the triplet accuracy of MP in all 10 model conditions (Figs. 3.3 and 3.4B, and Fig. C.5 in Appendix C. Overall, MV had lower error than MP (mean triplet error: 0.238 and 0.244, respectively), and the differences were statistically significant according to an analysis of variance (ANOVA) test comparing the two methods ($p < 10^{-5}$), and considering divergence from the clock or the outgroup distance as other independent variables (to be discussed in RQ2). However, averaged over all 7 conditions of D1 where outgroups were available, OG rooting was more accurate than MV rooting, a pattern that was not universal and will require a nuanced consideration of parameter effects (RQ2).

When we combine D1, D2, and D3 to get a heterogeneous dataset that ranges between 10 to 5000 taxa, a clear pattern emerges. While with smaller numbers of species, OG performs the best, when the number of taxa is increased to 1000 and beyond, MV gradually becomes the most accurate method (Fig. 3.3A). Increasing the number of taxa from 10 to 5,000 gradually increases the error for all methods, but OG is impacted more than MV (an increase from 0.1 triplet distance to 0.4 for OG, but from 0.2 to 0.35 for MV). MP is never the most accurate method but with trees of 5000 taxa, it is not worse than OG either. It is interesting to note that 2000 and 5000-taxon datasets, which have higher average tree height than 30-taxon datasets, have lower numbers of *true* gene trees where the outgroup species is not the gene tree outgroup (Fig. 3.1C). Thus, the sharp decrease in the accuracy of OG is not related to increased impacts of ILS and has to be attributed to increased error in the estimated gene trees. When we focus on our new datasets (D1 and D2), it becomes clear that improvements of MV over OG are the most pronounced with lower deviations from the clock (Fig 3.3B).

RQ2: Impact of error, clock, and outgroup distance

We focus our discussion on D1, but patterns on the D2 dataset are similar. On D1, we focus on the triplet error, but SPR distance gives similar results. See supplementary figures in Appendix C for details.

While MV is always at least as good as MP in our simulations, on D1, improvements of MV compared to MP are significantly impacted by both the level of divergence from the clock and the R/C ratio ($p = 0.002$ and $p < 10^{-5}$, respectively, according to the two-way ANOVA test). The improvements of MV over MP are higher when divergence from the clock is less and when the outgroup distance is smaller; the highest difference is for the case with no outgroup (Fig. 3.4A, Figs C.5 and C.6 in Appendix C).

The OG rooting is extremely accurate if true gene trees were known (Fig 3.4AB, and Figs. C.4 and C.5 in Appendix C); cases of error are limited to when the root is not very diverged from the ingroups (R/C$<$ 1). In contrast, MV and MP, while are better than OG with low divergence from the clock, can have very high error rate even on true gene trees if divergences from the clock are sufficiently large (Fig. 3.4B and Figs.C.4 and C.6 in Appendix C). For example, MV (MP) finds a root that on average has 25% (30%) normalized branch distance to the correct root (Fig. C.4 in Appendix C); i.e., the inferred root is away from the correct root by a quarter of the maximum tree height.

On estimated gene trees, however, the accuracy of OG rooting severely degrades. The delta triplet error (triplet error above ideal rooting) of OG is only slightly better than MV with various R/C ratios with medium divergence from the clock (Fig. 3.4A) and is worse than MV with low divergence from the clock (Fig. 3.4B); OG remains substantially more accurate than MV with high divergence from the clock. Confirming this pattern, considering all individual genes in all replicates, as gene tree error increases from 0% to approx. 50%, delta triplet error seems to increase for all methods but the increase is more pronounced for OG (Fig. 3.4C). Beyond 60% gene tree error (RF), delta triplet error actually goes down perhaps because even the ideal rooting has very high error, leaving little or no room for extra error due to rooting alone.

The delta triplet error of estimated gene trees rooted with OG reveals an interesting (U-shape) pattern. Choosing very small or very large R/C ratios (e.g., very close or distant outgroups) is not ideal (Fig. 3.4A). Instead, the best performance is obtained by R/C= 1. This ratio seems to give outgroups that are as close as possible to the ingroups to reduce LBA effects while remaining sufficiently long to reduce impacts of ILS.

RQ3: Species tree error

We focus on the average RF distance here; using RF distributions (Fig. C.3 in Appendix C) or average distances according to the MS metric (Table C in Appendix C) do not change any of our conclusions.

The average RF error of species trees run on estimated gene trees with inferred roots ranges between 9.1% and 9.5% (Table 3.1). STAR run on the true gene trees with the true root has an average RF error of 5.8%; thus, a substantial part of the species tree error can simply be attributed to ILS and lack of insufficient number of gene trees to find a perfect species tree. STAR run on the gene trees with ideal rooting has 8.6% RF error, which is a 48% increase from STAR run on true gene trees. These differences are statistically significant according to a two-way ANOVA test where the clock divergence parameter is the second independent variable. Therefore, the second substantial contributor to the species tree error is the gene tree estimation error.

Despite all the differences observed in the accuracy of rooting individual gene trees, we surprisingly found no clear evidence that the rooting error has a significant impact on the species tree accuracy. The RF error of STAR species trees run on estimated gene trees with ideal rooting (which uses the known true gene tree) was not significantly different from that of the STAR run on estimated gene trees rooted using OG or MV (Table 3.1). We also saw no statistically significant differences between species trees estimated from gene trees rooted using MV or OG. Thus, given estimated gene trees, which in our dataset had high rates of error (Fig. 3.1B), the delta error due to rooting inaccuracies does not seem to lead to much further reduction in accuracy. Consistent with this hypothesis, we also observed no statistically significant differences (Table 3.1) between STAR rooted using OG and NJst (which due to its strong parallels with STAR can be called unrooted STAR [8]).

On estimated gene trees, all rooting methods are negatively impacted by increased deviations from a strict clock (Tables 3.1 and C.3 in appendix C). The reduction may relate to increased unrooted gene tree estimation error with increased deviations (Fig. 3.1B); it may also be related to the fact that rooting becomes successively harder with stronger deviations from the strict clock (Fig. 3.4B).

3.3.2 Biological results

We tested MV rooting on an angiosperm dataset with 46 species and 310 genes [365], where the correct rooting has been a point of debate [318]. This dataset includes a single outgroup (*Selaginella*). We rooted each gene tree using both OG and MV, and compared gene trees with the

Methods compared	**p-value**		**Mean RF ST error**	
	method	**clock par.**	**1st method**	**2nd method**
STAR True vs STAR Ideal	$< 10^{-5}$	0.126	0.058	0.086
STAR Ideal vs STAR OG	0.551	0.0009	0.086	0.091
STAR Ideal vs STAR MV	0.144	$< 10^{-5}$	0.086	0.095
STAR OG vs STAR MV	0.476	$< 10^{-5}$	0.091	0.095
STAR OG vs NJst	0.623	0.00005	0.091	0.093

Table 3.1: Species tree estimation accuracy using rooted and unrooted gene trees. ANOVA tests were performed on the D1 (30-taxon) dataset for pairs of methods. RF error is used as the metric. The tests were performed on the subset of D1 where outgroup exists. For true gene trees, the true root is known. For estimated gene trees, the Ideal is the rooting position that minimizes triplet error to the true gene trees. p-values are shown for the significance of differences between the error of the two methods specified in each row, and for the differences in error among the three levels of clock divergence parameter, respectively.

published MP-EST species tree [365, 219] using the triplet distance *after* removing the outgroup from the gene trees and the species tree. The motivation for using this score is that we conjecture an incorrect rooting will tend to increase observed discordance of gene trees with the species tree. On this dataset, OG and MV essentially result in the same average triplet distance to the MP-EST species tree (19.2% for OG and 19.3% for MV) and their differences are not statistically significant (p-value=0.9). It's worth noting that excluding outgroups could have had reduced gene tree estimation error, and therefore, may have been a better approach overall.

3.4 Discussions

Our simulations made it clear that even if the outgroup distance to ingroups is twice as much as the most distant ingroups (i.e., R/C= 1), there can still be many *true* gene trees that fail to have the outgroup as sister to the ingroups (Fig. 3.1C). How often such cases of outgroup/ingroup mixing happens depends on the level of ILS, and by extension on the depth of the species tree and population size. Our 30-taxon dataset had numbers of generations that ranged between 407K and 9.1M generations in 90% of replicates; thus, our trees range between relatively shallow to moderately deep. Overall congruence of gene trees with the species tree, as measured by the quartet score, was high (> 0.8) for 43% of replicates, and was moderate (0.6 – 0.8) for another 34%. Thus, despite having realistic conditions, we observe outgroups mixed with ingroups in true gene trees.

Making outgroups maximally distant from ingroups, however, won't solve the problem. As Rosenfeld *et al.* have pointed out [281], making the outgroups distant can lead to random assignment of outgroups in the gene trees, thereby increasing the apparent discordance. In agreement with their results, and much of the literature, we found that very distant outgroups, while placed as desired in the true gene trees, can lead to increased overall error (Fig. 3.4A). There is a trade-off between making the outgroup closer to ingroups to minimize LBA and making it more distant to reduce ILS; in our 30-taxon dataset and under our conditions of simulations, the optimal setting was R/C=1, corresponding to outgroups that are twice as distant from any of the ingroups as the most two divergent ingroups. The exact optimal value, however, likely depends on the exact parameters of a biological dataset and the choice of R/C= 1 cannot be blindly prescribed.

Increased divergence from the clock substantially increased unrooted gene tree estimation error (Fig. 3.1B), but impacted the accuracy of rooting only when MP or MV were used (Fig. 3.4B). The strong dependence of gene tree estimation on clock assumptions leads us to suggest that simulations of the MSC process should always include conditions where the strict clock are violated. Many methods are proved consistent and tested empirically only under the strict clock assumption, a situation that we hope our results will change. New simulation tools such as SimPhy make it easy to simulate datasets that deviate from the strict clock assumption.

A surprising result of our simulation studies was that while gene tree rooting error was generally high, we could not detect a significant impact on the species tree. Two explanations have to be considered. It could be that in general the impact of rooting error on species tree estimation is minimal. On the other hand, the lack of power to detect significant impact may be limited to specifics of our simulation procedure. Several important parameters of the simulation may have reduced the effect of rooting error on species tree estimation error. We always had five hundred genes, which is relatively high considering that we only had 30 ingroup species. Impact of rooting error for datasets with more species and/or fewer genes may be different, a problem that we did not get to address here because of computational limitations. Moreover, we conjecture that at least part of the reason for this lack of observed impact is that our datasets had high levels of gene tree estimation error even for the unrooted tree. It is conceivable that the impact from mis-rooting is drawn out by the impact of topological error and is hard to detect with a datasets of 100 replicates, simulated with heterogeneous parameters drawn from wide parameter distributions. We note that our simulation setup was designed mainly to address the question of gene tree rooting error and to enable a comparison between our new MV and existing MP and OG rooting methods. Moreover, we focused only on comparing NJst and STAR because of their deep mathematical connection; our current study cannot be generalized to other methods such as ASTRAL and MP-EST (which can in principle be altered to take as input both rooted and unrooted trees). Thus, while our results are suggestive that there may be considerable robustness to gene tree rooting error at least among some methods, to arrive at a more nuanced understanding of impacts of rooting, simulation setups designed directly to answer this questions will be needed in future.

Several other limitations of our study should be noted. In our simulations, we always included only one outgroup (a limitation of SimPhy), but the impact of selecting multiple outgroups will be important to examine. We inferred gene trees under the exact model of sequence evolution that generated the data, but the impact of factors such as LBA are known to be exacerbated by model misspecification. Our deviations from the clock were random and did not depend on time. Finally, more realistic models of change in evolutionary tempo may result in more systematic biases and different conclusions.

3.5 Conclusion

We introduce a new method for rooting phylogenetic trees, which relies on minimizing the variance of the root to tip distances. The method can be efficiently implemented in an algorithm that scales linearly with increased number of species and runs in less than a minute for datasets of up to 200,000 leaves. Our new approach is more accurate than the traditional midpoint rooting and its relative accuracy compared to the dominant method of outgroup rooting depends on the number of species; with very large trees, minimizing root to tip variance outperforms outgroup rooting whereas for small and moderate size datasets outgroups are more accurate. Regardless of the relative accuracy of methods, we showed that rooting gene trees is challenging because deviations from a strict clock make it hard for automatic methods to find the correct root, while gene tree discordance makes outgroup rooting unreliable. However, within the limitations of our study, we detected no significant impact due to gene tree error on the accuracy of the species tree accuracy for datasets with large numbers of gene trees, many of them inferred from datasets with low phylogenetic signal. We leave a more nuanced consideration of impacts of incorrect rooting on species tree error to future research.

3.6 Acknowledgements

Chapter 3, in full, is a reprint of the material as it appears in “Minimum variance rooting of phylogenetic trees and implications for species tree reconstruction”. Mai, Uyen, Erfan Sayyari, and Siavash Mirarab. PloS one 12, no. 8 (2017): e0182238. The dissertation author was the primary investigator and first author of this paper.

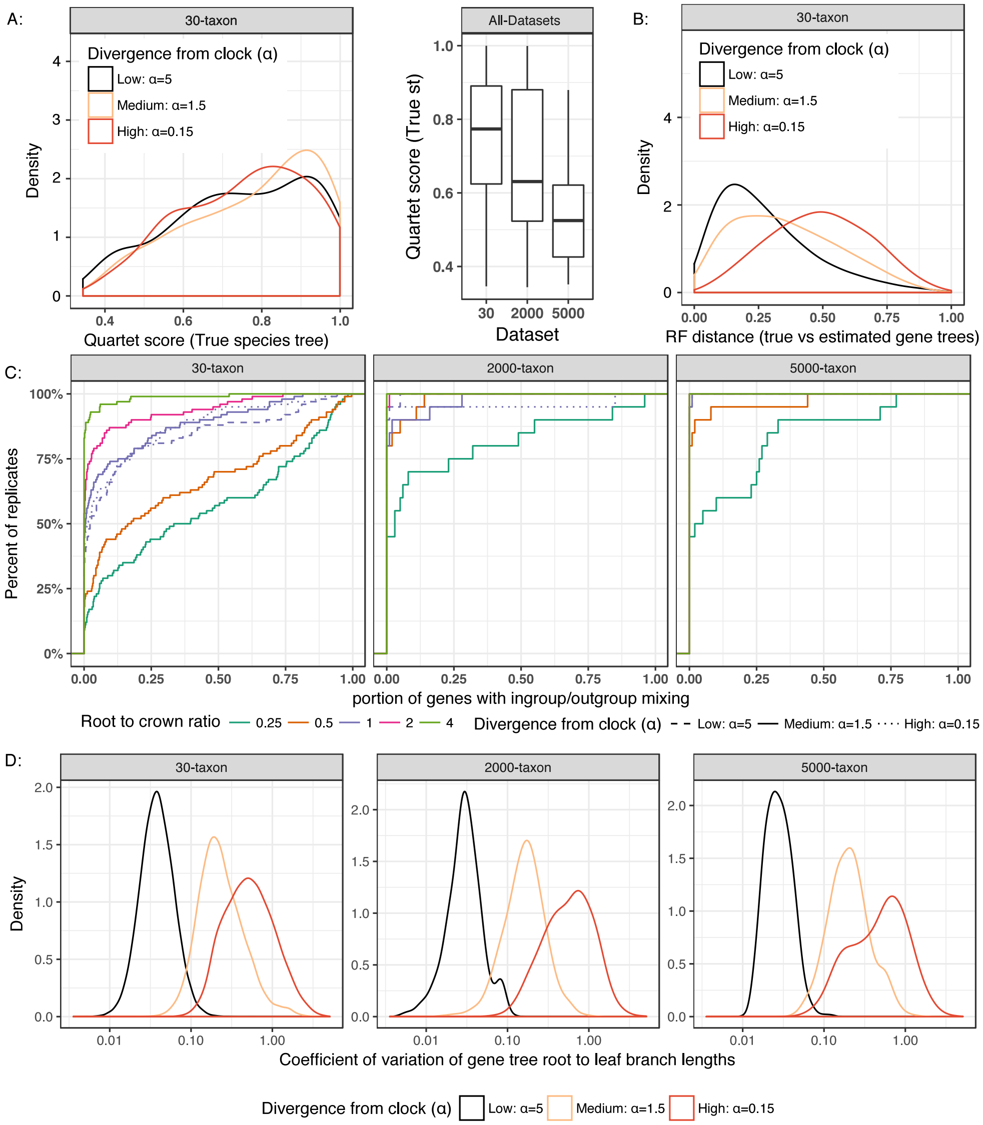


Figure 3.1: Properties of simulated datasets D1 and D2. A: The level of ILS, measured by the quartet score of true species tree with respect to true gene trees with R/C= 1 for (left) the D1 dataset, broken down by the clock divergence parameter and (right) both D1 and D2 datasets. B: gene tree estimation error, measured as the normalized Robinson-Foulds (RF) distance [278] between true and estimated gene trees for the D1 dataset with R/C= 1 and varying clock divergence parameters. C: The empirical cumulative distribution for the proportion of true gene trees where the outgroup species is not an outgroup; thus, each point (x,y) on a line indicates that y out of 100 replicates had *at most* $x \times 500$ true gene trees where the species tree outgroup was not the gene tree outgroup. Boxes correspond to the three datasets with different sizes. D: The ratio between standard deviation to mean (i.e., coefficient of variation) of root to leaf distances of gene tree branches, as an empirical measure of divergence from the clock; 0 corresponds to strict molecular clock and higher values correspond increased divergence (the x axis is in the log scale). See Appendix C for model conditions not shown here.

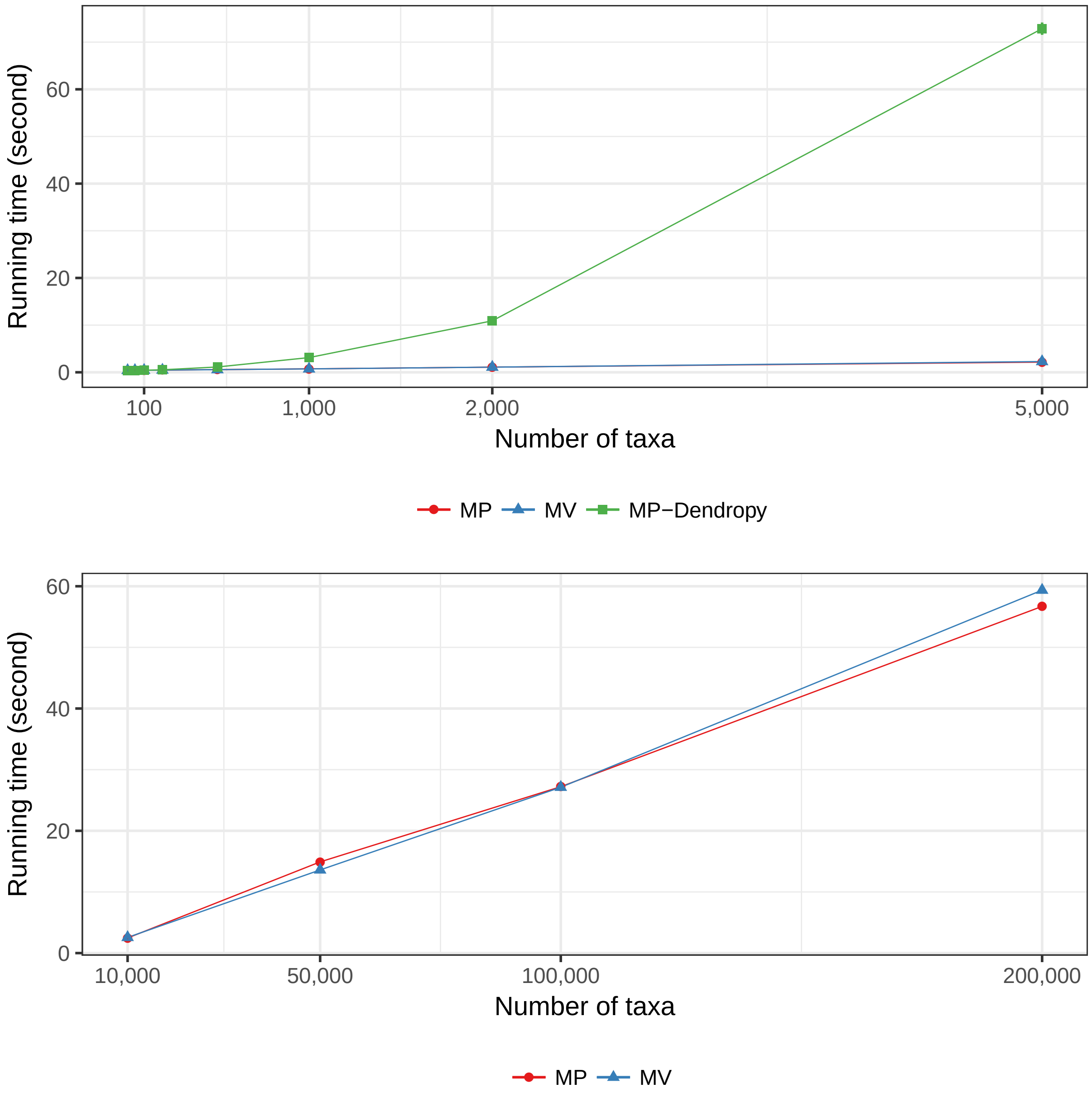


Figure 3.2: Running time of MP and MV. Left: comparison of our implementation of MV/MP with the implementation of MP in Dendropy, which employs a quadratic algorithm, on datasets D1, D2, and D3 with up to 5,000 leaves; Right: Linear time scaling of our implementation, tested on the RNASim dataset with up to 200,000 leaves.

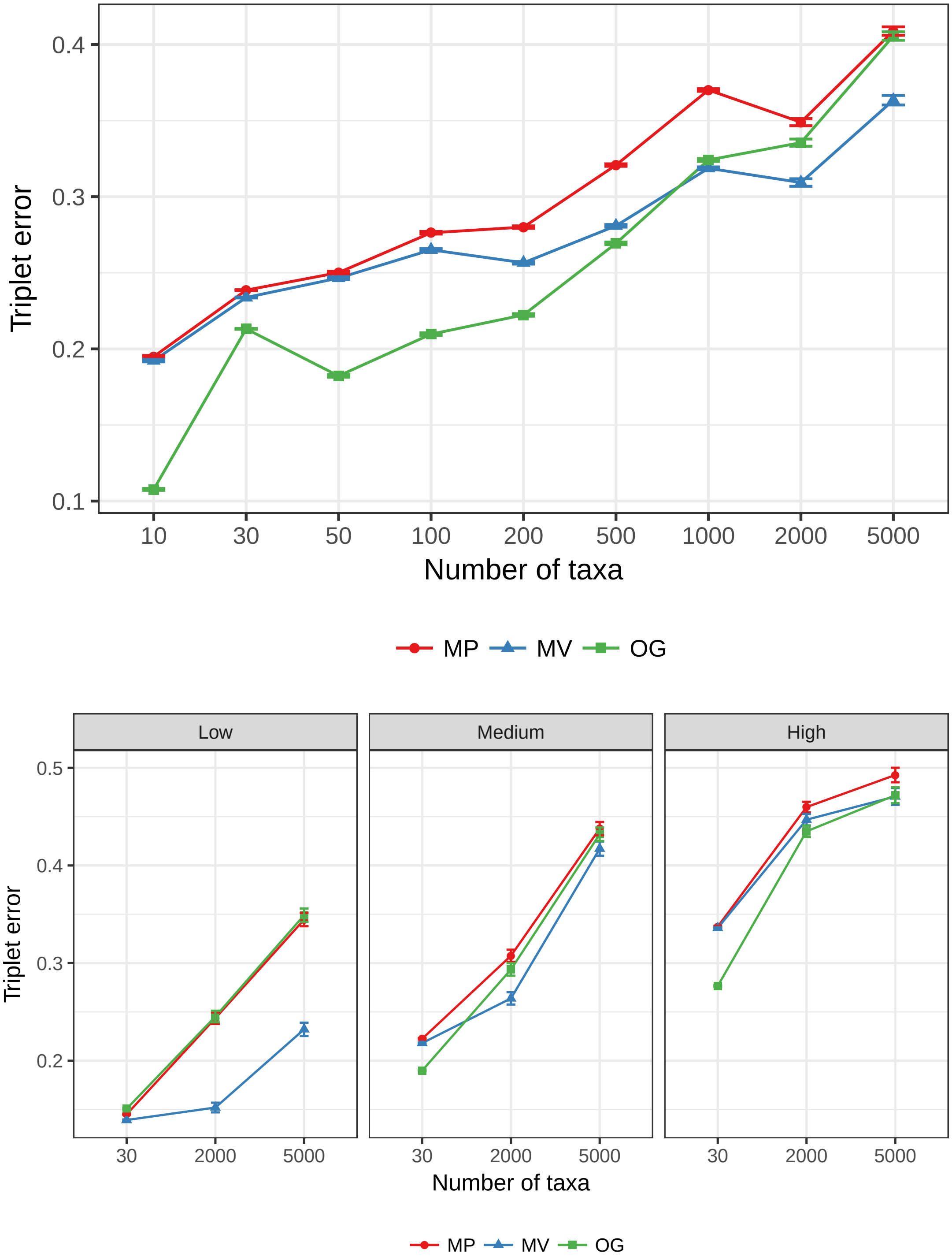


Figure 3.3: Absolute triplet distance as a function of the number of taxa. Top: Results from D1, D2, and D3 are combined in one figure; 30 on the x-axis corresponds to D1, 2000 and 5000 to D2, and the remaining cases to D3. For D1 and D2, we fixed R/C= 1 and the clock divergence parameter to medium to best match the conditions of D3. Bottom: Results for D1 and D2 with R/C= 1 and difference levels of clock divergence.

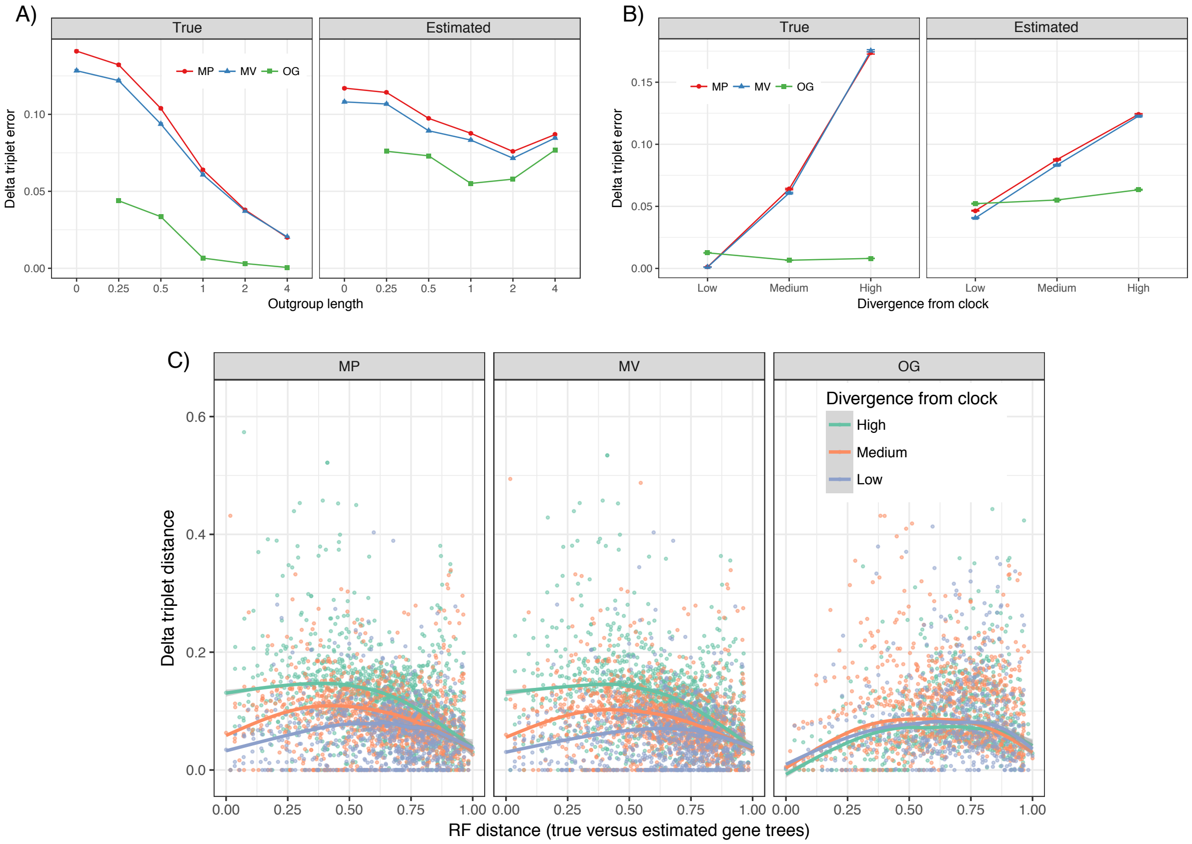


Figure 3.4: Rooting error above ideal rooting on 30-taxon dataset. Top: delta triplet error with both true and estimated gene trees for (A) medium divergence from the clock and varying R/C ratios and (B) R/C=1 and varying levels of divergence from the clock. C: Delta triplet error versus gene tree estimation error, measured by RF distance, shown for high, medium, and low divergence from the clock; each point is an average of all gene trees in all replicates that had an identical RF gene tree error. A loess regression is fitted to the data using R.

Chapter 4

Log Transformation Improves Dating of Phylogenies

Phylogenetic trees inferred from sequence data often have branch lengths measured in the expected number of substitutions and therefore, do not have divergence times estimated. These trees give an incomplete view of evolutionary histories since many applications of phylogenies require time trees. Many methods have been developed to convert the inferred branch lengths from substitution unit to time unit using calibration points, but none is universally accepted as they are challenged in both scalability and accuracy under complex models. Here, we introduce a new method that formulates dating as a non-convex optimization problem where the variance of log-transformed rate multipliers are minimized across the tree. On simulated and real data, we show that our method, wLogDate, is often more accurate than alternatives and is more robust to various model assumptions.

4.1 Introduction

Phylogenetic inference from sequence data does not reveal divergence time (i.e. exact timing of evolutionary events) unless paired with external timing information. Under standard models of sequence evolution, the evolutionary processes, including sequence divergence, are fully determined by the product of the absolute time and mutation rates in a non-identifiable form. Thus, these models measure branch lengths in the unit of expected numbers of mutations per site (since standard models like GTR [336] only allow substitutions, focusing on these models, we use *substitutions* and *mutations* interchangeably throughout this paper). Nevertheless, knowing divergence times is crucial for understanding evolutionary processes [130, 99] and is a fundamental need in many clinical applications of phylogenetics and phylodynamics [347]. A commonly used approach first infers a phylogeny with branch lengths in the unit of substitution per site and then dates the phylogeny by translating branch lengths from substitution unit to time unit; co-estimation of topology and dates is also possible [77] though its merits have been debated [351].

The fundamental challenge in dating is to find a way to factorize the number of substitutions into the product of the evolutionary rate and time. A common mechanism allowing this translation is to impose soft or hard constraints on the timing of *some* nodes of the tree, leaving the divergence times of the remaining nodes to be inferred based on the constrained nodes. Timing information is often in one of two forms: calibration points obtained from the geological record [168] and imposed on either internal nodes or tips that represent fossils [74], or tip sampling times for fast-evolving viruses and bacteria. The constraints still leave us with a need to extrapolate from observed times for a few nodes to the remaining nodes, a challenging task that requires a mathematical approach. Obtaining accurate timing information and formulating the right method of extrapolation are both challenging [283].

Many computational methods for dating phylogenies are available [283, 171], and a main point of differentiation between these methods is the clock model they assume [291]. Some methods rely on a strict molecular clock [384] where rates are effectively assumed to be constant [177, 304]. However, empirical evidence has now made it clear that rates can vary substantially, and ignoring these changes can lead to incorrect dating [39, 170]. Consequently, there have been many attempts to *relax* the molecular clock and allow variations in rates. A main challenge in relaxing the clock is the need for a model of rates, and it is not clear what model should be preferred. As a result, many methods for dating using relaxed molecular clocks have been developed. Some of these methods allow rates to be drawn independently from a stationary distribution [77, 346, 6] while others model the evolution of rates with time [143] or allow correlated rates across branches [339, 166, 292, 185, 79, 331, 313]. Despite these developments, strict molecular clocks continue to be used, especially in the context of intraspecific evolution where there is an expectation of relatively uniform rates [42].

Another distinction between methods is the use of explicit models [290]. Many dating methods use a parametric statistical model and formulate dating as estimating parameters in a maximum likelihood (ML) or Bayesian inference framework [177, 77, 346, 340]. Another family of methods [292, 331] formulate dating as optimization problems, including distance-based optimization [367, 366], that avoid computing likelihood under an explicit statistical model. When the assumed parametric model is close to the reality, we expect parametric methods to perform well. However, these methods can be sensitive to model deviations, a problem that may be avoided by methods that avoid using specific models.

In this paper, we introduce LogDate, a new method of dating rooted phylogenies that allows variations in rates but without modeling rates using specific distributions. We define mutation rates necessary to compute time unit branch lengths as the product of a single global rate and a set of rate multipliers, one per branch. We seek to find the overall rate and all rate multipliers such that the log-transformed rate multipliers have the minimum variance. This formulation gives us a constrained optimization problem, which although not convex, can be solved in a scalable fashion using the standard approaches such as sequential least squares programming. While formulation of dating as an optimization problem is not new [340, 177], here we introduce log-transformation of the rate multipliers, which as we will show, results in more accurate dates. Our observation is in line with a recent change to RelTime [332] where the switch from arithmetic means to geometric means (between rates of sister lineages) has improved accuracy. In extensive simulation studies and three biological data sets, we show that a weighted version of LogDate, namely wLogDate, has higher accuracy in inferring node ages compared to alternative methods, including some that rely on time-consuming Bayesian inference. While wLogDate can date trees using both sampling times for leaves (e.g., in viral evolution) or estimated time of ancestors, most of our results are focused on cases with sampling times at the tips of the tree.

4.2 Methods

4.2.1 Definitions and notations

For a rooted binary tree T with n leaves, we give each node a unique index in $[0,\dots,2n-2]$. By convention, the root is always assigned 0, the other internal nodes are arbitrarily assigned indices in the range $[1,\dots,n-1]$, and the leaves are arbitrarily assigned indices in the range $[n-1,\dots,2n-2]$. In the rest of this paper, we will refer to any node by its index. If a node i is not the root node, we let $par(i)$ denote the parent of i and if i is not a leaf, we let $c_l(i)$ and $c_r(i)$ denote the left and right children of i, respectively. We refer to the edge connecting $par(i)$ and i as e_i.

We can measure each edge e_i of T in either time unit or substitution unit. Let t_i denote the divergence time of node i, i.e. the time when species i diverged into $c_l(i)$ and $c_r(i)$. Then for any node i other than the root, $\tau_i = t_i - t_{par(i)}$ is the length of the edge e_i in time unit. We measure divergence time of a node with respect to a fixed reference point in the past (i.e., time increases forward). Thus, we enforce $t_i > t_{par(i)}$ for all i. Let μ_i be the substitution rate (per sequence site per time unit) on branch e_i, then the expected number of substitutions per sequence site is $b_i = \mu_i\tau_i$. Let $\tau = [\tau_1, \dots, \tau_{2n-2}]$ and $\mathbf{b} = [b_1, \dots, b_{2n-2}]$.

From sequence data, $\mathbf{b}$ can be inferred using standard methods such as maximum parsimony [97], minimum evolution [284], neighbor-joining [286, 105], and maximum-likelihood (ML) [91, 118, 231]. Note that inferred trees need to be rooted subsequently using an outgroup (that can be removed) or automatic methods such as midpoint or minimum variance rooting [201]. We let $\hat{b}_i$ denote the estimate of b_i by an inference method and let $\hat{\mathbf{b}} = [\hat{b}_1, \dots, \hat{b}_{2n-1}]$.

In this paper, we are interested in computing τ from $\hat{\mathbf{b}}$. The computation of τ from $\hat{\mathbf{b}}$ is complicated by two factors: (1) the possibility of change among rates, and (2) deviations of the inferred edge length $\hat{b}_i$ from the true value b_i.

To better describe the mathematical formulation of the optimization problem, we first do the following change of variables. Assuming the mutation rates on the branches are distributed around a global rate μ, we define $\nu_i = \frac{\mu\tau_i}{\hat{b}_i}$. Let $\mathbf{x} = [\nu_1, \dots, \nu_{2n-2}, \mu]$; our goal of finding τ is identical to finding $\mathbf{x}$.

4.2.2 Dating as a constrained optimization problem

We formulate dating as an optimization problem on $2n-1$ variables $\mathbf{x} = [\nu_1, \dots, \nu_{2n-2}, \mu]$, subject to the linear constraints defined by calibration points and/or sampling times. Many existing methods, including LF [177] and LSD [340], can be described in this framework, with the choice of the objective function distinguishing them from each other. We start by describing the setup of the constraints enforced by a set of calibration points/sampling times, and show that they can all be written as linear equations on $\mathbf{x}$. We then give the formulation of both LF and LSD in this framework and use their formulation to motivate our own new approach. Finally, we describe strategies to solve the wLogDate optimization problem.

Linear constraints Ψ from sampling times

For any pair of nodes (i,j) (where each of i and j can either be a leaf or an internal node) with enforced divergence times (t_i,t_j), the following constraint $\psi(i,j)$ must be satisfied

$$\psi(i,j): \mu(t_j - t_i) = \sum_{k \in P(m,j)} \nu_k \hat{b}_k - \sum_{k \in P(i,m)} \nu_k \hat{b}_k \tag{4.1}$$

where m is the LCA of i and j and $P(m,j)$ and $P(i,m)$ are the paths connecting m to j and i to m, respectively. Thus, given k time points, $k(k-1)/2$ constraints must hold. However, only $k-1$ of these constraints imply all others, as we show below.

Let t_0 be the *unknown* divergence time at the root of the tree. For k calibration points $t_1,\dots,t_k$, we can setup k constraints of the form:

$$C_i : \mu(t_i - t_0) = \sum_{k \in P(0,i)} \nu_k \hat{b}_k, \tag{4.2}$$

where node 0 is the root and $P(0,i)$ is the path from the root to node i. For any pair (i,j), the linear constraint given in Eq. 4.1 can be derived by subtracting C_i from C_j side by side. Also, we can remove t_0 from the set of constraints by subtracting C_1 from all other constraints $C_2,\dots,C_k$. This gives us the final $k-1$ linear constraints on $\mathbf{x}$, which we denote as Ψ. We can build Ψ using Algorithm 1 (Supplementary material).

Optimization Criteria

Since $\nu_i = \frac{\mu\tau_i}{\hat{b}_i}$, the distribution of ν_i is influenced by both the distribution of the rates (μ_i) and the distribution of $\hat{b}_i$ around b_i. In traditional strict-clock models [384], a constant rate is assumed throughout the tree ($\forall_i \mu_i = \mu$). Under this model, the distribution of ν_i is determined by deviations of $\hat{b}_i$ from b_i.

[177] (LF) modeled the number of *observed* substitutions *per sequence site* on a branch i

by a Poisson distribution with mean $\lambda = \mu\tau_i$ and treated $s\hat{b}_i$ as if they were the total number of *observed* substitutions; as such, they assume $s\hat{b}_i \sim \mathit{Poisson}(s\mu\tau_i)$, where s is the sequence length. Therefore, by changing variable, we can write the log-likelihood function as:

$$\sum_{i=1}^{2n-2} \big(s\hat{b}_i \log(s\hat{b}_i) - \log((s\hat{b}_i)!)\big) + \sum_{i=1}^{2n-2} s\hat{b}_i \left(\log \nu_i - \nu_i\right).$$

Given s and $\hat{b}_i$, LF finds $\mathbf{x}$ that maximizes the log-likelihood function and subject to the constraints Ψ. As such,

$$\mathbf{x}_P^* = \arg\min_{\mathbf{x}} \sum_{i=1}^{2n-2} \hat{b}_i \big(\nu_i - \log \nu_i\big) \text{ subject to } \Psi. \tag{4.3}$$

[340] assume $\hat{b}_i$ follows a Gaussian model: $\hat{b}_i \sim \mathit{Gaussian}(\mu\tau_i, \sigma_i^2)$ and assume the variance is approximated by $\frac{\hat{b}_i}{s}$ (the method includes smoothing strategies omitted here). Then, the negative log likelihood function can be written as:

$$\sum_{i=1}^{2n-2} \frac{(\hat{b}_i - \mu\tau_i)^2}{\sigma_i^2} \approx \sum_{i=1}^{2n-2} \frac{s}{\hat{b}_i}(\hat{b}_i - \mu\tau_i)^2 = \sum_{i=1}^{2n-2} s\hat{b}_i (1 - \nu_i)^2.$$

Thus, the ML estimate can be formulated as:

$$\mathbf{x}_G^* = \arg\min_{\mathbf{x}} \sum_{i=1}^{2n-2} \hat{b}_i (1 - \nu_i)^2 \quad \text{subject to } \Psi. \tag{4.4}$$

Both LF and LSD have convex formulations. [177] proved that their negative log-likelihood function is convex and thus the local minimum is also the global minimum. Our constraint-based formulation of LF also can be easily proved convex by showing its Hessian matrix is positive definite. [340] pointed out their objective function is a weighted least squares. Using our formulation, we also see that Eq. 4.4 together with the calibration constraints form a standard convex quadratic optimization problem which has a unique analytical solution.

4.2.3 LogDate Method

Motivation

LF only seeks to model the errors in $\hat{\mathbf{b}}$ and ignore true rate heterogeneity. Strict-clock assumption is now believed to be unrealistic in many settings [265, 301, 135], motivating relaxed clocks, typically by assuming that μ_is are drawn i.i.d. from some distribution [77, 346, 6]. Most methods rely on presumed parametric distributions (typically, LogNormal, Exponential, or Gamma) and estimate parameters using ML [346], MAP [6], or MCMC [77, 78]. The LSD method, which like LF directly models errors in $\hat{\mathbf{b}}$, is additionally justified under a normally-distributed clock model. Choices of specific distributions in these methods are not motivated by the knowledge that real data follow them exactly (for example, the Normal distribution has to be misspecified as mutation rates cannot be negative).

Our goal is to avoid explicit parameter inference under a model of rate multipliers. Instead, we follow the assumption shared by existing methods like LSD and LF: we assume that given two solutions of $\mathbf{x}$ both satisfying the calibration constraints, the solution with less variability in ν_i values is preferable. Thus, we prefer solutions that minimize deviations from a strict clock while allowing deviations. A natural way to minimize deviations from the clock is to minimize the variance of $\frac{\tau_i}{\hat{b}_i}$. This can be achieved by finding μ and all ν_i such that ν_i is centered at 1 and $\sum_{i=1}^{2n-2}(\nu_i - 1)^2$ is minimized. Interestingly, the ML function used by LSD (Eq. 4.4) is a weighted version of this approach.

The minimum variance principle results in a fundamental asymmetry: multiplying or dividing the rate of a branch by the same factor are penalized differently (Fig 4.1a). For example, the penalty for $\nu_i = 4$ is more than ten times larger than $\nu_i = 1/4$. The LF model is more symmetrical than LSD but remains asymmetrical (Fig 4.1a). This asymmetry results from the asymmetric distribution of the Poisson distribution around its mean, especially for small mean, in log scale (Fig 4.1b). Because of this asymmetry, methods like LSD and LF judge a very small

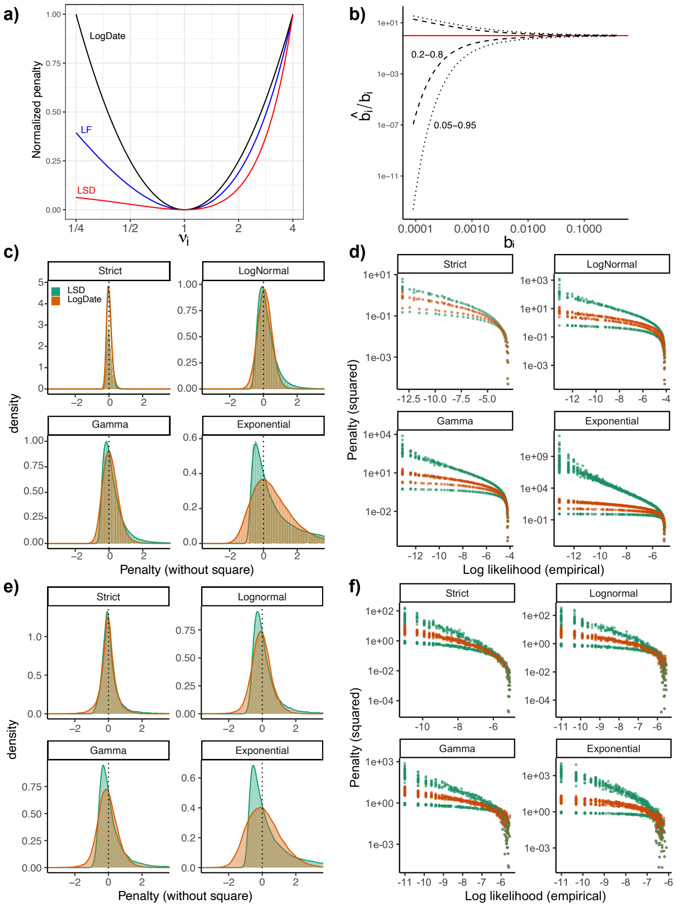


Figure 4.1: **(a)** The penalty associated to multiplying a single edge i with multiplier ν_i in LSD, LF, and LogDate approaches, as shown in Equations 4.3, 4.4, and 4.5. To allow comparison, we normalize the penalty to be zero at $\nu = 1$ and to be 1 at $\nu = 4$. **(b)** The confidence interval of the ratio between estimated and true branch length using the Poisson model. For this purpose of this exposition, we assume that the estimated branch length equals the number of substitutions occurring on the branch and follows a Poisson distribution (i.e., JC69 model), divided by sequence length. With these assumptions, the CI for estimate length $\hat{b}_i$ is between $^1/_2\chi^2_{2sb_i}$ and $^1/_2\chi^2_{2sb_i+2}$; we draw the CI for $^{\alpha}/_2 = 0.05$ and $^{\alpha}/_2 = 0.2$ to get 0.2–0.8 and 0.05–0.095 intervals for $0.0001 \leq b_i \leq 0.4$. **(c)** and **(e)**: Density and histograms of penalty terms (without square) used by LSD ($^{\mu\tau_i}/_{\hat{b}_i} - 1$) and LogDate ($\log {^{\mu\tau_i}/_{\hat{b}_i}}$) under different clock models. (c) Fixing $\mu\tau_i = 0.1$, we draw 500000 rate multipliers (r_i) from LogNormal, Gamma, or Exponential distributions with mean 1 and variance 0.16 for LogNormal and Gamma. For strict clock, $r_i = 1$. We then draw estimated branch length for each replicate i from the Normal distributed with mean $b_i = r_i\mu\tau_i$ and variance $^{b_i}/_s$ for $s = 200$. (e) The branch lengths are estimated from the sequences using PhyML from simulated sequences of [340], as explained in the text. Parameters of rate multiplier distributions match part (c). We omit extremely short branches (< 0.001) for better visualization. **(d)** and **(f)**: The penalty of LSD and LogDate versus the empirical log-likelihood of estimated length for the models described in (c) and (e), respectively. To compute the empirical likelihood, we divide estimated branch lengths into small bins and the empirical likelihood of each bin is estimated as the frequency of the data assigned to it. See Appendix D Fig. D.2 for extended results.

$\hat{b}_i/b_i$ to be within the realm of possible outcomes, and thus penalize $\nu_i < 1$ multipliers less heavily than $\nu_i > 1$.

Our method is based on a principle, which we call the *symmetry of ratios*: the penalty for multiplying a branch by a factor of ν should be no different than dividing the branch by ν. Note that this assertion is only applicable to true variations of the mutation rate (i.e, ignoring branch length estimation error). We further motivate this principle with more probabilistic arguments below, but here we make the following case. If one considers the distribution of rate *multipliers* for various branches, absent of an explicit model, it is reasonable to assume that compared to an overall rate, branches rates are as likely to increase by a factor of ν as they are to decrease by a factor of ν. When this statement is true, we shall prefer a method that penalizes ν and $1/\nu$ identically. To ensure the symmetry of ratios, we propose taking the logarithm of the multipliers ν_i before minimizing their variance. Minimizing the variance of the rates in log-scale is the essence of our method. It achieves the symmetry, and, as we show below, a better correspondence between penalty and data likelihood.

Log-transformation has long been used to reduce data skewness before applying linear regression [325, 164, 368, 55]. In molecular dating, it can be argued that log-transformation is implicit in the new version of RelTime [332] where the geometric means between sister lineages replaced the arithmetic means in its predecessor. The improvement in the accuracy of RelTime encourages a wider use of log-transformation in molecular dating. Note that log-transforming the rate multipliers before minimizing their least squares penalty is identical to applying linear least squares after log-transformation of both time and the number of substitutions. In other fields, log-transformation has been used to make the least-squares method more robust to highly skewed distributions [211, 2].

LogDate optimization function

We formulate the LogDate problem as follows. Given $\hat{\mathbf{b}}$ and the set of calibration constraints described earlier, we seek to find

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \sum_{i=1}^{2n-1} \log^2(\nu_i) \quad \text{subject to } \Psi. \tag{4.5}$$

This objective function satisfies the symmetry of ratio property (Fig. 4.1a). Since ν_i values are multipliers of rates around μ, if we assume μ is the mean rate, the LogDate problem is equivalent to minimizing the variance of the log-transformed rate multipliers (around their mean 1). The objective function only depends on ν_i; however, note that μ is still included in the constraints and therefore is part of the optimization problem. This setting reduces the complexity of the objective function and speeds up the numerical search for the optimal solution. Since the values of ν_i close to 1 are preferred in Eq. 4.5, the optimal solution would push μ to the mean rate.

Justification as a relaxed-clock model

After log-transformation, LogDate, similar to LSD, constructs the objective function using the least squares principle (for ease of exposition, here we discuss ordinary least-squares without weights). We can rewrite the objective function of LSD as $\sum_i (\frac{\mu\tau_i}{\hat{b}_i} - 1)^2$ and that of LogDate as $\sum_i (\log \frac{\mu\tau_i}{\hat{b}_i})^2$ and see that both seek to find a global rate μ and the time τ_i for each branch to minimize the total deviations of the estimated branches from $\mu\tau_i$. This observation may motivate viewing both LSD and LogDate as strict-clock methods. However, the following result justifies viewing LogDate as a relaxed clock method.

We can prove that if the mutation rates μ_i are drawn i.i.d. from a LogNormal distribution with any parameters with mode μ and the branches are estimated without error (i.e. $\hat{b}_i = b_i$ for all i), then ν_i follows a LogNormal distribution with mode 1 and the LogDate optimization problem is equivalent to finding ν that have maximum joint probability, subject to the constraints. The proof is given in Claim 1 (Supplementary materials).

Justification for symmetry of ratios

Having shown that LogDate has a justification under the LogNormal distribution, we now compare LogDate and LSD objective functions in a wider range of clock models. Recall that the objective functions of LSD and LogDate are the sum-of-squares of their penalty terms, which are $\frac{\mu\tau_i}{\hat{b}_i} - 1$ for LSD and $\log\frac{\mu\tau_i}{\hat{b}_i}$ for LogDate.

Following the likelihood principle, an ideal objective function must assign equal penalties to data values that are equally likely to occur. Therefore, for an ideal objective function written as sum-of-squares of the penalty terms, the probability distribution of its penalty terms (before square) under the model that generates the data must be symmetric around 0 (because of the square). The true distribution of our penalty terms is a function of both clock rate variations and branch length estimation error. While no objective function is ideal for all compound models of rates and estimation error, a robust objective function should remain close to symmetric and maintain a low skewness under a wide range of models. We now present several theoretical and empirical results comparing LogDate and LSD in terms of skewness of distributions of their penalty terms.

First, consider a relaxed clock model of the rates and assume no branch estimation error (i.e., $\hat{b}_i = \mu_i\tau_i$). If μ_i follows a LogNormal distribution parameterized by θ and σ then it is easy to see that $\frac{\mu\tau_i}{\hat{b}_i} = \frac{\mu}{\mu_i}$ (penalty of LSD) also follow a LogNormal distribution and the skewness depends on σ. In contrast, the $\log\frac{\mu\tau_i}{\hat{b}_i}$ (penalty of LogDate) follows a Normal distribution, which has skewness 0, and for which least square estimation is the maximum likelihood estimator. Thus, as stated before, log-transforming is the ML solution if rate multipliers are log-normally distributed.

Now assume μ_i follows a Gamma distribution with mean μ. Then ${}^{\mu\tau_i}/_{\hat{b}_i} = {}^{\mu}/_{\mu_i}$ follows an Inverse Gamma distribution while its log-transformation follows a Log-Gamma distribution. We can analytically compute the skewness of the penalty terms of LSD and LogDate and compare them (see Supplementary materials for the equations). As shown in Fig. D.1 (Appendix D), the skewness of LSD is much higher than that of LogDate, especially for higher variance of the gamma rates. Higher skewness of penalty terms violates the likelihood principle mentioned before. Thus, for the two models where we could compute analytical formulas for skewness, we have grounds to prefer LogDate.

Next, we consider the compound impacts of branch length estimation error and rate variation, and we study the question in two ways. One approach is to measure the combined effect of error and true variation by simulating sequence data and measuring $\hat{b}_i$ for known b_i empirically; here, we use simulations by [340] with 1000 sites and PhyML-inferred trees (details will be provided in the Experiments section). The other approach is modelling the compound effect. While it is hard to know generally how estimated branch length is distributed around its expected value, here, we can follow [340] and assume $\hat{b}_i \sim \mathcal{N}(b_i, \frac{b_i}{s})$. The other challenge is that the compound distribution of estimation error and rate multipliers is hard to compute analytically. However, we can easily generate a very large number of samples from compound distributions and analyze the empirical distribution to approximate the true distribution.

Inspecting the empirical density of the penalty terms of LSD and LogDate across different clock models result in consistent patterns using both approaches, modeling the compound distribution (Fig. 4.1c) and using simulated sequence data (Fig. 4.1e). Across three models of rates, Exponential, LogNormal, and Gamma, the distributions of the LogDate penalty terms are always more symmetric than that of LSD. Results are similar for other rate models such as Log-Uniform and are further amplified when the variance is increased (Fig. D.2a, Appendix D).

To further explore that relationship between the likelihood and the penalty assigned by LogDate and LSD, we plot the penalty (with square terms) versus the empirical log likelihood of the rate multipliers (Figs. 4.1d and 4.1f and S2b in Supplementary Materials). Ideally, increasing likelihood should monotonically decrease penalty, and points with similar likelihood should have similar penalties. In both modelled and simulated branch lengths and across models, LSD assigns two sets of widely different penalties (one for increased and one for decreased rates) to data with similar likelihood. LogDate, while far from perfect, is much closer to the ideal mapping between likelihood and penalty. Also, for LogNormal with *median* rate multipliers set to 1, we empirically observe a perfectly monotonic relationship between the penalty and likelihood (Fig. D.2b in Appendix D), as theory suggested.

wLogDate optimization function

The simple LogDate formulation, however, has a limitation: by allowing rates to vary freely in a multiplicative way, it fails to deal with the varied levels of relative branch error; i.e., the ratio of the estimated branch length to the true branch length ($\hat{b}_i/b_i$). As $\hat{b}_i$ is estimated from the sequences, the error of $\hat{b}_i$ is directly related to the variations in the number of substitutions occurred along the branch b_i. Let us assume sequences follow the [157] model, and le N_i be the total number of substitutions occurred along branch i on a sequence with length s. Under Juke-Cantor model, we have $N_i \sim Poisson(s\mu\tau_i)$ and therefore, $var(N_i) = s\mu\tau_i$. Therefore, the variance of the *expected number of substitutions* around the true branch length is $var(\frac{N_i}{sb_i}) = \frac{s\mu\tau_i}{s^2b_i^2} = \frac{1}{b_is}$. As Figure 4.1b shows, when b_i is small, $\frac{N_i}{s}$ can easily vary by several orders of magnitude around b_i. Furthermore, the distribution is not symmetric: drawing values several factors smaller than the mean is more likely than drawing values above the mean by the same factor. These analyses predict that the distribution of $\frac{\hat{b}_i}{b_i}$ depends strongly on b_i - with smaller b_i giving higher variance - and is not symmetric.

The variances of the relative error $\frac{\hat{b}_i}{b_i}$ is difficult to compute analytically due to the involvement of the sequence substitution model and the method to estimate $\hat{b}_i$, which are both unknown. Therefore, we instead use empirical analyses of the estimated branch lengths by PhyML to demonstrate our arguments. Consistent with our prediction, Figures S7 a and c illustrate that the relative error $\frac{\hat{b}_i}{b_i}$ varies more in small branches and the distribution is not symmetric. These properties of the branch length estimates are not modeled in our LogDate formulation and we seek to incorporate them in a refined version of LogDate which will be described below.

Since the true branch length b_i is unknown, a common practice is to use the estimated $\hat{b}_i$ in place of b_i to estimate its variance as $\frac{1}{\hat{b}_i s}$. This explains why both LF and LSD objective functions (Eqs. 4.3 and 4.4) have a weight of $\hat{b}_i$ for each term of ν_i. Following the same strategy, we propose weighting each $\log^2(\nu_i)$ term in a way that reduces the contribution of short branches to the total penalty, and thus allows more deviations in the log space if the branch is small (and is thus subject to higher error). Since we log-transform ν_i and pursue a model-free approach, explicitly computing the weights to cancel out the variations of relative error among the branches is challenging. However, since the weights should reflect the variance of $\frac{\hat{b}_i}{b_i}$ (logarithmic scale), they should monotonically increase with $\hat{b}_i$ (Fig. 4.1b) to allow more variance for the *relative* errors in short branches than in long branches. We use $\sqrt{\hat{b}_i}$ as weights, a selection driven by simplicity and empirical performance (shown in a later section).

The shortest branches require even more care. When the branch is very short, for a limited-size alignment, the evolution produces zero mutations with high probability. For these no-event branches, tree estimation tools report arbitrary small lengths (see Fig. D.7 in Appendix D), rendering $\hat{b}_i$ values meaningless for very small branches. To deal with this challenge, the r8s's implementation of LF [293] collapses all branches with length $\hat{b}_i < {}^{1}/_{s}$. [340] proposed adding a smoothing constant ${}^{c}/_{s}$ to each $\hat{b}_i$ to estimate the variance of $\hat{b}_i$, where c is a parameter that the user can tune. Following a similar strategy, we propose adding a small constant $\tilde{b}$ to each $\hat{b}_i$. We choose $\tilde{b}$ to be the maximum branch length that produces no substitutions with probability at least $1-\alpha$ for $\alpha \in [0,1]$. Recall that N is the total number of *actual* substitutions on a branch. Under the [157] model, it is easy to show that $\arg\max_{\tilde{b}} Pr(N=0|b=\tilde{b}) \geq 1-\alpha = -\frac{1}{s}\log(1-\alpha)$. We choose this value as $\tilde{b}$ and set $\alpha = 0.01$ by default. Thus, we define the weighted LogDate (wLogDate) as follows:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \sum_{i=1}^{2n-1} \sqrt{\tilde{b}+\hat{b}_i}\log^2(\nu_i) \qquad \text{subject to } \Psi. \tag{4.6}$$

Solving the optimization problem

Both LogDate and wLogDate problems (Eq. 4.5 and Eq. 4.6) are non-convex, and hence solving them is non-trivial. The problem is convex if $0 \leq \nu_i \leq e$. For small clock deviation and small estimation error in $\hat{b}_i$, the ν_i values should be small so that the problem becomes convex with one local minimum. However, as $\nu_i \leq e$ is not guaranteed, we have to rely on gradient-based numerical methods to search for multiple local minima and select the best solution we can find. To search for local minima, we use the Scipy solver with `trust-constr` [175] method. To help the solver work efficiently, we incorporate three techniques that we next describe.

Computing Jacobian and Hessian matrices analytically helps speedup the search. By taking the partial derivative of each ν_i, we can compute the Jacobian, J, of Eq. 4.6. Also, since Eq. 4.6 is separable, its Hessian H is a $(2n-2) \times (2n-2)$ diagonal matrix. Simple derivations give us:

$$\begin{aligned} J = &\left[2\sqrt{\tilde{b}+\hat{b}_1}\frac{\log\nu_1}{\nu_1}, \ldots, 2\sqrt{\tilde{b}+\hat{b}_{2n-2}}\frac{\log\nu_{2n-2}}{\nu_{2n-2}}\right]^T \\ \text{and} \quad H_{ii} = &2\sqrt{\tilde{b}+\hat{b}_i}\frac{1-\log\nu_i}{\nu_i^2} \; . \end{aligned}$$

Sparse matrix representation further saves space and computational time. The Hessian matrix is diagonal, allowing us to store only the diagonal elements. In addition, the constraint matrix defined by Ψ is highly sparse. If all sampling times are given at the leaves, the number of non-zero elements in our $(n-1) \times (2n-1)$ matrix is $O(n\log n)$ (Claim 3; Supplementary materials). If the tree is either caterpillar or balanced, the number of non-zeroes reduced to $\Theta(n)$. Thus, we use sparse matrix representation implemented in the Scipy package. This significantly reduces the running time of LogDate.

Starting from multiple feasible initial points is necessary given that our optimization problem is non-convex. Providing initial points that are feasible (i.e. satisfied the calibration constraints) helps the SciPy solver work efficiently. We designed a heuristic strategy to find multiple initial points given sampling times $t_1,\dots,t_n$ of all the leaves (as is common in phylodynamics).

We first describe the process to get a single initial point. We compute the root age t_0 and μ using root-to-tip regression (RTT) [304]. Next, we scale all branches of T to conform with Ψ as follow: let $m = \arg\min_i t_i$ (breaking ties arbitrarily). Let $d(r,i)$ denote the distance from the root r to node i and $P(r,m)$ denote the path from r to m. For each node i in $P(r,m)$, we set $\tau_i = \hat{b}_i(t_m - t_0)/d(r,m)$. Then going upward from m to r following $P(m,r)$, for each edge (i,j) we compute $t_j = t_i - \tau_i$ and recursively apply the process on the clade i. At the root, we set t_m to the second oldest (second minimum) sampling time and repeat the process on a new path until all leaves are processed. Since RTT gives us μ, to find ν we simply set $\nu_i = {}^{\mu\tau_i}/_{\hat{b}_i}$.

To find multiple initial points, we repeatedly apply RTT to a set of randomly selected clades of T and scale each clade using the aforementioned strategy. Specifically, we randomly select a set S of some internal nodes in the tree and add the root to S. Then, by a post-order traversal, we visit each $u \in S$ and date the clade u using the scaling strategy described above. We then remove the entire clade u from the tree but keep the node u as a leaf (note that the age of u is already computed) and repeat the process for the next node in S. The root will be the last node to be visited. After visiting the root, we have all the τ_i for all i. After having all the branches in time unit, we find $\mathbf{x}$ to minimize either Eq. 4.5 or Eq. 4.6, depending on whether LogDate or wLogDate is chosen. In a tree of n leaves, we have $2^{(n-1)} - 1$ ways to select the initial non-empty set S, giving us enough room for randomization.

Computing confidence interval

With the ability of wLogDate to work on any combination of sampling times/calibration points on both leaves and internal nodes (as long as at least two time points are provided), we design a method to estimate the confidence intervals for the estimates of wLogDate. We subsample the sampling times/calibration points given to us repeatedly to create *N* replicate datasets (where *N* is 100 by default, but can be adjusted). Note our subsampling is not exactly a bootstrapping procedure as node sampling times cannot be resampled with replacement. We then compute the time tree for each replicate to obtain *N* different estimates for the divergence time of each node, from which we can compute their confidence intervals (95% as default). This sampling would work best when we have a fairly large number of calibration points, which is the case in phylodynamic settings where all (or nearly all) sampling times for the leaves are given, or in large phylogenies where abundant calibration points can be obtained from fossils. Although we refer to the resulting intervals as confidence intervals, it is important to recognize that the resampling procedure is not strictly justified via bootstrap theory because subsampling is necessarily without replacement and sampled nodes are not independent of each other.

4.2.4 Experiments on simulated data

Phylodynamics setting

[340] simulated a dataset of HIV *env* gene. Their time trees were generated based on a birth-death model with periodic sampling times. There are four tree models, namely D995_11_10 (M1), D995_3_25 (M2), D750_11_10 (M3), and D750_3_25 (M4), each of which has 100 replicates for a total of 400 different tree topologies. M1 and M2 simulate intra-host HIV evolution and are ladder-like while M3 and M4 simulate inter-host evolution and are balanced. Also, M4 has much higher root-to-tip distance (mean: 57) compared to M1–M3 (22, 33, and 29). Starting from conditions simulated by [340], we use the provided time tree to simulate the

clock deviations. Using an uncorrelated model of the rates, we draw each rate from one of three different distributions, each of which is centered at the value $\mu = 0.006$ as in [340]. Thus, we set each μ_i to $x_i\mu$ where x_i is drawn from one of three distributions: LogNormal (mean:1.0, std: 0.4), Gamma ($\alpha = \beta = 6.05$), and Exponential ($\lambda = 1$). Sequences of length 1000 were simulated for each of the model conditions using SeqGen [270] under the same settings as [340].

Calibrations on autocorrelated rate model

We used the software NELSI and the same protocol as in [133] to simulate a dataset where the rates are autocorrelated. The dataset has 10 replicates, each contains 50 taxa. The time trees were generated under Birth-death model and the rate heterogeneity through time is modeled by the autocorrelation model ([166]) with the initial rate set to 0.01 and the autocorrelated parameter set to 0.3. DNA sequences (1000 bases) were generated under Jukes-Cantor model. We used PhyML [118] to estimate the branch lengths in substitution unit from the simulated sequences while keeping the true topology. These trees are the inputs to wLogDate, RelTime, LF, and DAMBE [366] to infer time trees.

4.2.5 Real biological data

H1N1 2009 pandemic We re-analyze the H1N1 biological data provided by [340] which includes 892

H1N1pdm09 sequences collected worldwide between 13 March 2009 and 9 June 2011. We reuse the estimated PhyML [118] trees, 100 bootstrap replicates, and all the results of the dating methods other than LogDate that are provided by [340].

San Diego HIV We study a dataset of 926 HIV-1 subtype B *pol* sequences obtained in San Diego between 1996 and 2018 as part of the PIRC study. We use IQTree [231] to infer a tree under the GTR+Γ model, root the tree on 22 outgroups, then remove the outgroups. Because of the size, we could not run BEAST.

West African Ebola epidemic We study the dataset of Zaire Ebola virus from Africa, which includes 1,610 near-full length genomes sampled between 17 March 2014 and 24 October 2015. The data was collected and analyzed by [80] using BEAST and re-analyzed by [346] using IQTree to estimate the ML tree and *treedater* to infer node ages. We run LSD, LF, and wLogDate on the IQTree from [346] and use the BEAST trees from [80], which include 1000 sampled trees (BEAST-1000) and the Maximum clade credibility tree (BEAST-MCC). To root the IQTree, we search for the rooting position that minimizes the triplet distance [288] between the IQTree and the BEAST-MCC tree.

Methods Compared

For the phylodynamics data, we compared wLogDate to three other methods: LSD [340], LF [177], and BEAST [78]. For all methods, we fixed the true rooted tree topology and only inferred branch lengths. For LSD, LF, and wLogDate, we used phyML [118] to estimate the branch lengths in substitution unit from sequence alignments and used each of them to infer the time tree. LSD was run in the same settings as the QPD* mode described in the original paper [340]. LF was run using the implementation in r8s [293]. wLogDate was run with 10 feasible starting points. For the Bayesian method BEAST, we also fixed the true rooted tree topology and only inferred node ages. Following [340], we ran BEAST using HKY+Γ8 and coalescent with constant population size tree prior. We used two clock models on the rate parameter: the strict-clock (i.e. fixed rate) model and the LogNormal model. For the strict-clock prior, we set clock rate prior to a uniform distribution between 0 and 1. For the LogNormal prior, we set the ucld.mean prior to a uniform distribution between 0 and 1, and ucld.stdev prior to an exponential distribution with parameter $^1/_3$ (default). We always set the length of the MCMC chain to 10^7 generations, burn-in to 10%, and sampling to every 10^4 generations (identical to [340]).

For the autocorrelated rate model, we compared wLogDate to LF and RelTime [332], which is one of the state-of-the-art model-free dating methods. We randomly chose subsets of the internal nodes (10% on average) as calibration points and created 20 tests for each of the 10 replicates (for a total of 200 tests). We also compared wLogDate to DAMBE using this dataset. Because DAMBE can only be run in interactive mode where each calibration point has to be manually placed onto the tree, we could not run DAMBE on the 200 tests with hundreds of calibration points in total. Therefore, we instead ran DAMBE only once on each of the 10 trees and infer a unit time tree for each of them (i.e. calibrate the root to be at 1 unit time backward) and compared the results to that of wLogDate. DAMBE does not accept identical sequences so we removed identical sequences from the simulated alignments and trees before running DAMBE and ran wLogDate using these reduced trees to have a fair comparison.

Evaluation Criteria

On the simulated phylodynamics dataset where the ground truth is known, we compare the *accuracy* of the methods using several metrics. We compute the root-mean-square error (RMSE) of the true and estimated vector of the divergence times (τ) and normalize it by tree height. We also rank methods by RMSE rounded to two decimal digits (to avoid different ranks when errors are similar). In addition, we examine the inferred divergence time of the Most Recent Common Ancestor (tMRCA) and mutation rate. The comparison of methods mostly focuses on point-estimates of these parameters and the accuracy of the estimates (as opposed to their variance). In one analysis, we also compare the confidence intervals produced by wLogDate and BEAST on one model condition (M3 with LogNormal rate distribution). Finally, we examine the correlation between variance of the error in wLogDate and divergence times and branch lengths.

On the simulated data with autocorrelated rate, we show the distributions of the divergence times estimated by wLogDate, LF, and RelTime and report the RMSE normalized by tree height for each replicate. To compare to DAMBE in inferring unit time trees, we report the average relative error of the inferred to the true divergence times. After removing identical sequences, there are 438 internal nodes in total across the 10 tree replicates. For each internal nodes, we compute the relative error of its divergence time inferred by either DAMBE or wLogDate to its true divergence time in the normalized true time tree, which is $\frac{|\hat{t_i}-t_i|}{t_i}$ where $\hat{t_i}$ and t_i are the inferred and true divergence times of node i, respectively. We report the average relative error per tree replicate and the average of all 438 nodes for DAMBE and wLogDate.

On real data, we show lineage-through-time (LTT) plots [229], which trace the number of lineages at any point in time and compare tMRCA times to the values reported in the literature. We also compare the runtime of wLogDate to all other methods in all analyses.

4.3 Results

4.3.1 Simulated data for phylodynamics

We first evaluate the convergence of the ScipPy solver across 10 starting points (Fig. D.3a in Appendix D). LogDate and wLogDate converge to a stable result after 50–200 iterations, depending on the model condition. Convergence seems easier when rates are Gamma or LogDate and harder when the rates are Exponential. Next, to control for the effect of the starting points on the accuracy of our method, we compare the error of these starting points with the wLogDate optimal point (Fig. D.3b in Appendix D). In all model conditions, the optimal point shows dramatic improvement in accuracy compared to the starting point. We then compare different weighting strategies for LogDate (Table D.4 in Appendix D).

In all model conditions, the weighting $\sqrt{\hat{b}_i+\tilde{b}}$, is one of the two best, so it is chosen as the default weighting for wLogDate. Moreover, wLogDate is never worse than LogDate, and under exponential clock models, appropriate weighting results in dramatic improvements (Table D.4 in Appendix D).

Next, we study the properties of wLogDate estimates in relation to: (1) the age of the

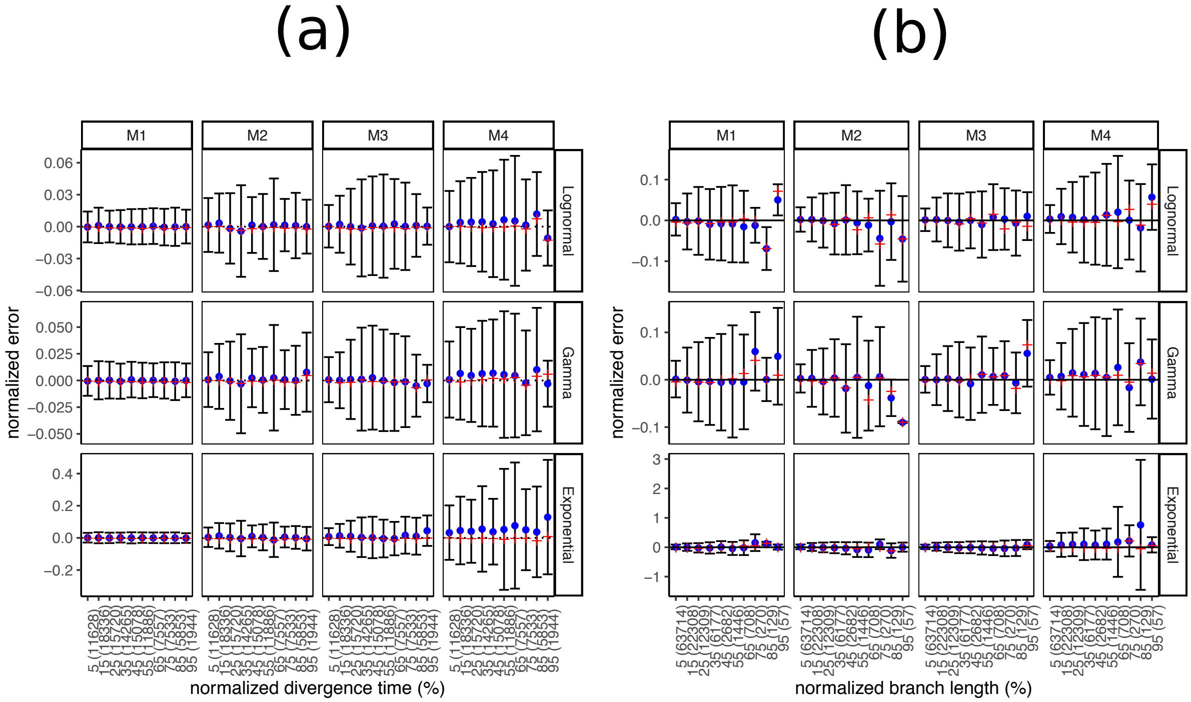


Figure 4.2: Analyses of wLogDate on inferring branch lengths on simulated data. (a) error normalized by tree height versus divergence time (i.e. the time of the midpoint of each branch); both axes are normalized by the tree height. (b) error versus branch length (in time unit); both axes are normalized by the maximum branch length. For both (a) and (b), the x-axis is discretized into 10 bins of equal size. We label the bins by their median values, relative to either the tree height for (a) or the maximum branch length for (b). We also show the number of points in each bin in parentheses. Note the small number of points in the final bins in panel (b). For each bin, the blue dot represents the mean, the red cross represents the median, and the bar represents one standard deviations around the mean.

node (Fig. 4.2a), (2) the length of the true branch in time unit (Fig. 4.2b), and (3) the error of the branch lengths (in substitution unit) estimated by PhyML (Fig. D.6). Overall, we do not observe a substantial change in the mean estimation error of wLogDate as the node age and the branch length change. The variance, however, can vary with node ages (Figure 4.2a), especially in M3 and M4 model conditions. Moreover, longer branches have a tendency to have higher variance in absolute terms (Fig. 4.2b). However, note that the relative error (i.e., log-odds error) dramatically

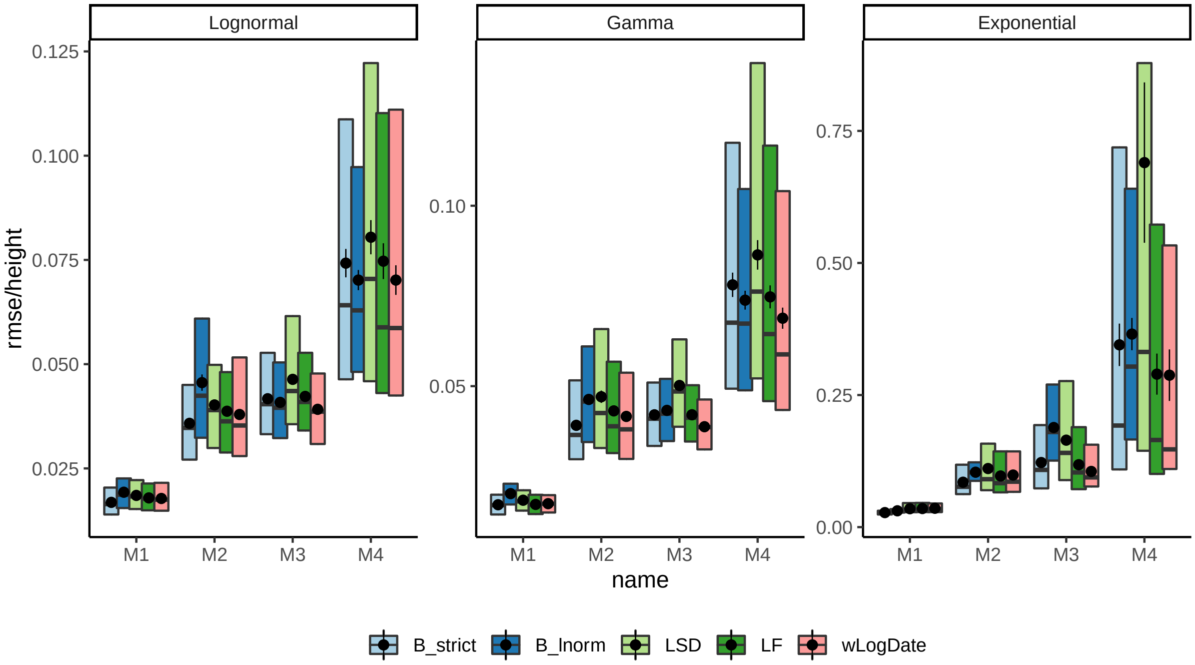


Figure 4.3: Distributions of RMSE normalized by the tree height for internal node ages inferred by all methods on model trees M1–M4, each with clock models Lognorm, Gamma, and Exponential. Boxes show median, 10% and 90% quantiles; dots and error bars show mean and standard error (100 replicates).

reduces as branches become longer (Fig. D.6 in Appendix D).

In studying the effect of the error in branch length estimation, we see that wLogDate can underestimate the branch time if the branch length in substitution unit is extremely underestimated (Fig. D.6a in Appendix D). In some cases wLogDate under-estimates branch times by two order of magnitude or more; all of these cases correspond to super-short branches with substitution unit branch length under-estimated by three or four orders of magnitude . As mentioned previously, extremely short estimated branch lengths are often the zero-event branches (Fig. D.7 in Appendix D), which are unavoidable for short sequences.

We next compare wLogDate to alternative methods, namely LF, LSD, and BEAST with strict-clock and Lognormal clock. Measured by RMSE, the accuracy of all methods varies substantially across model trees (M1 – M4) and models of rate variation (Fig. 4.3). Comparing methods, for many conditions, wLogDate has the lowest error, and in many others, it is ranked

model	Clock model	B_lnorm	B_strict	LF	LSD	wLogDate
M4	LogNormal	**1**	3	4	5	**1**
	Gamma	2	4	3	5	**1**
	Exponential	4	3	2	5	**1**
M3	LogNormal	2	3	3	5	**1**
	Gamma	4	2	2	5	**1**
	Exponential	5	3	2	4	**1**
M2	LogNormal	5	**1**	3	4	2
	Gamma	4	**1**	3	5	2
	Exponential	4	**1**	2	5	3
M1	LogNormal	4	**1**	2	4	2
	Gamma	5	**1**	**1**	4	**1**
	Exponential	2	**1**	3	3	5
average rank		3.5	2	2.5	4.5	**1.75**

Table 4.1: Ranking of the dating methods under different model conditions. For each model condition, the average RMSE of all internal node ages is computed and ranked among the dating methods (rounded to two decimal digits). The best method is shown in bold.

second best (Table 4.1). Across all conditions, wLogDate has a mean rank of 1.75, followed by BEAST with strict clock with a mean rank 2; mean normalized RMSE of wLogDate, LF, BEAST-strict, BEAST-LogNormal, and LSD are 0.072, 0.074, 0.077, 0.087, and 0.116, respectively. Interestingly, in contrast to wLogDate, LSD seems to often underestimate branch times for many short branches even when they are estimated relatively accurately in substitution units (Fig. D.6b in Appendix D).

For all methods, errors are an order of magnitude smaller for the LogNormal and Gamma models of rate variations compared to the Exponential model. In terms of trees, M4, which simulates inter-host evolution and high the largest height, presents the most challenging case for all methods. Interestingly, wLogDate has the best accuracy under all parameters of M4 tree and also all parameters of M3 (thus, both inter-host conditions). On M1, all methods have very low error and perform similarly (Fig. 4.3).

Among other methods, results are consistent with the literature. Despite its conceptual similarity to wLogDate, LSD has the worst accuracy. On M1 and M2, LSD is competitive with other methods; however, on M3 and M4, it has a much higher error, especially with the Exponential model of rate variation. With the LogNormal clock model, BEAST-LogNormal is better than BEAST-strict only for M4 but not for M1–M3; in fact, BEAST-LogNormal has the highest error for the M2 condition. This result is surprising given the correct model specification. Nevertheless, BEAST-LogNormal is competitive only under the LogNormal model of rate variation and is one of the two worst methods elsewhere. Thus, BEAST-LogNormal is sensitive to model misspecification. In contrast, BEAST-strict is less sensitive to the model of rate variation and ranks among the top three in most cases. In particular, BEAST-strict is always the best method for intra-host ladder-like trees M1 and M2.

Accuracy of tMRCA follows similar patterns (Fig. 4.4). Again, the Exponential rate variation model is the most difficult case for all methods, resulting in biased results and highly variable error rates for most methods. In all conditions of M3 and M4, wLogDate has the best accuracy and improves on the second best method by 9 – 66% (Table 4.2). For M1 and M2, BEAST-strict is often the best method. The mean tMRCA error of wLogDate across all conditions is 4.83 (years), which is substantially better than the second best method, BEAST-strict (6.21).

In terms of the mutation rate, the distinction between methods is less pronounced (Table S1). wLogDate is the best method jointly with the two strict clock models BEAST-strict and LF. Overall, even though LF and wLogDate tend to over-estimate mutation rates, both have less biased results compared to other methods (Fig. 4.4). LSD and BEAST-LogNormal have the highest errors; depending on the condition, each can overestimate or underestimate the rate but LSD tends to underestimate while BEAST-LogNormal tends to overestimate. On M1, wLogDate and LF have a clear advantage over BEAST-strict, which tends to over-estimate the rate. On M2, the three methods have similar accuracy. For M3 and M4, BEAST-strict under-estimates the rate under the Exponential model of rate variation, and wLogDate and LF are closer to the true value.

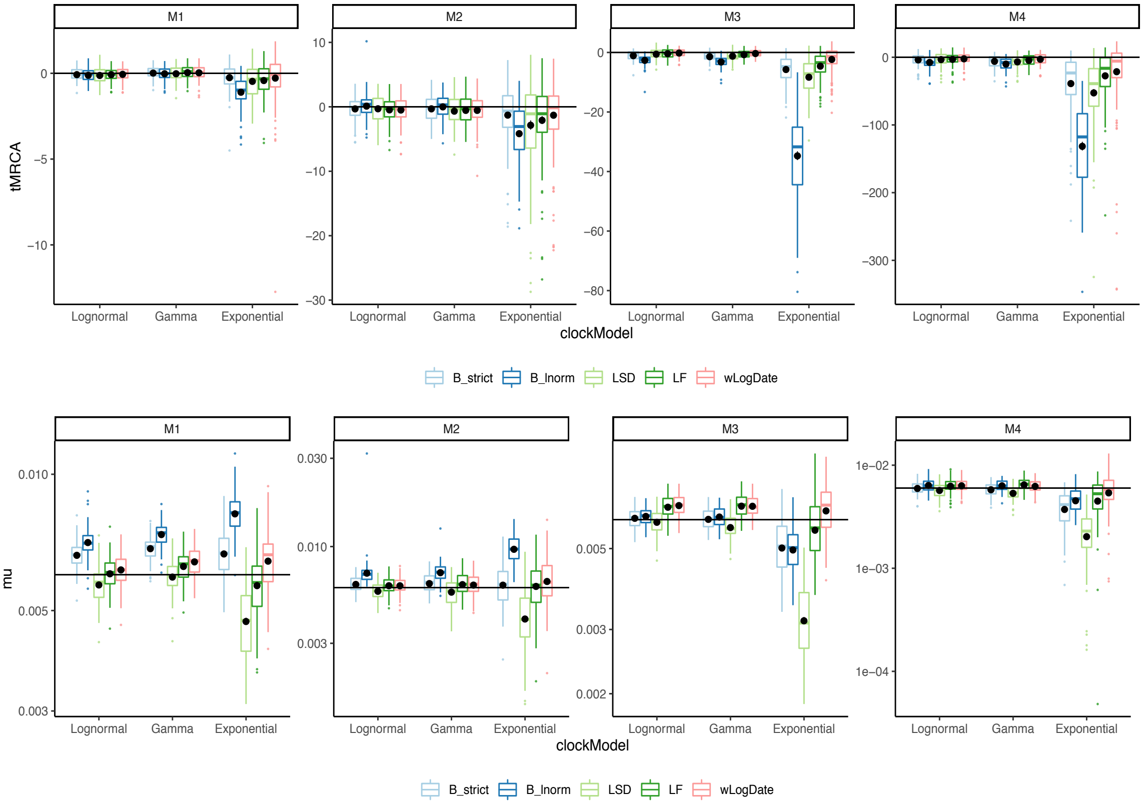


Figure 4.4: The inferred (top) tMRCA and (bottom) expected mutation rate on different tree models and clock models. Distributions are over 100 replicates. The solid horizontal lines indicate the true mutation rate and tMRCA. Each black is the average of the inferred values for each method under each model condition. We remove 6 outlier data points (2 LF, 1 LSD, 2 BEAST-LogNormal, 1 BEAST-Strict) with exceptional incorrect tMRCA (< -350) in the M4/Exponential model.

For all methods, M4 is the most challenging case.

We also compare confidence intervals obtained from wLogDate and BEAST (Fig.4.5). Although wLogDate intervals are on average 2.7 times larger than BEAST, 33% and 12% of the true values fall outside the 95% confidence interval for BEAST and wLogDate, respectively. Thus, while both methods under-estimate the confidence interval range, wLogDate, with its larger intervals, is closer to capturing the true value in its confidence interval at the desired level.

Finally, we compared all methods in terms of their running time (Table S2). LSD and LF are the fastest methods in all conditions, always taking tens of seconds (less than a minute) on

Tree	Clock Model	B_strict	B_lnorm	LF	LSD	RTT	wLogDate
M4	Lognormal	6.99	9.50	6.66	7.38	9.28	**6.11** (9% ↓)
	Gamma	7.83	10.48	7.02	8.48	8.24	**6.28** (12% ↓)
	Exponential	43.5	140.9	116.2	62.2	**31.5**	32.5 (3% ↑)
M3	Lognormal	1.37	2.60	1.21	1.39	1.46	**1.03** (17% ↓)
	Gamma	1.60	3.14	1.23	1.67	1.42	**0.97** (27% ↓)
	Exponential	5.76	34.67	4.87	8.35	3.39	**2.94** (66% ↓)
M2	Lognormal	**1.40**	1.41	1.50	1.63	2.19	1.47 (5% ↑)
	Gamma	1.54	**1.44**	1.75	1.92	2.56	1.66 (15% ↑)
	Exponential	**3.39**	4.59	4.28	5.27	5.23	3.72 (10% ↑)
M1	Lognormal	**0.28**	**0.28**	0.30	0.37	0.78	0.30 (7% ↑)
	Gamma	**0.27**	0.29	0.32	0.35	0.80	0.30 (11% ↑)
	Exponential	**0.60**	1.11	0.79	0.82	1.37	0.69 (15% ↑)
Average		6.21	17.54	12.17	8.13	5.68	**4.83**

Table 4.2: Mean absolute error of the inferred tMRCA of BEAST_strict, BEAST_lognorm, LF, LSD, RTT, and wLogDate. For wLogDate, parenthetically, we compare it with the best (↑) or second best (↓) method for each condition. We show percent improvement by wLogDate, as measured by the increase in the error of the second best method (wLogDate or the alternative) divided by the error of the best method.

these data. The running time of wLogDate depends on the model condition and can be an order of magnitude higher for Exponential rates than the other two models of rate variation. Nevertheless, wLogDate finishes on average in half a minute to 12 minutes, depending on the model condition. In contrast, BEAST took close to one hour with strict clock and close to two hours with the LogNormal model (and even more if run with longer chains; see Table S5 in Supplementary Materials.

4.3.2 Simulated data with autocorrelated rate

In simulations with the autocorrelated rate model, we compare wLogDate to LF and RelTime (Fig. 4.6 and Table S7) and wLogDate to DAMBE (Table D8 in Appendix D). The distribution of the estimated divergence time of uncalibrated internal nodes does not show any sign of biased in divergence time estimation for either method. All methods seem to give less

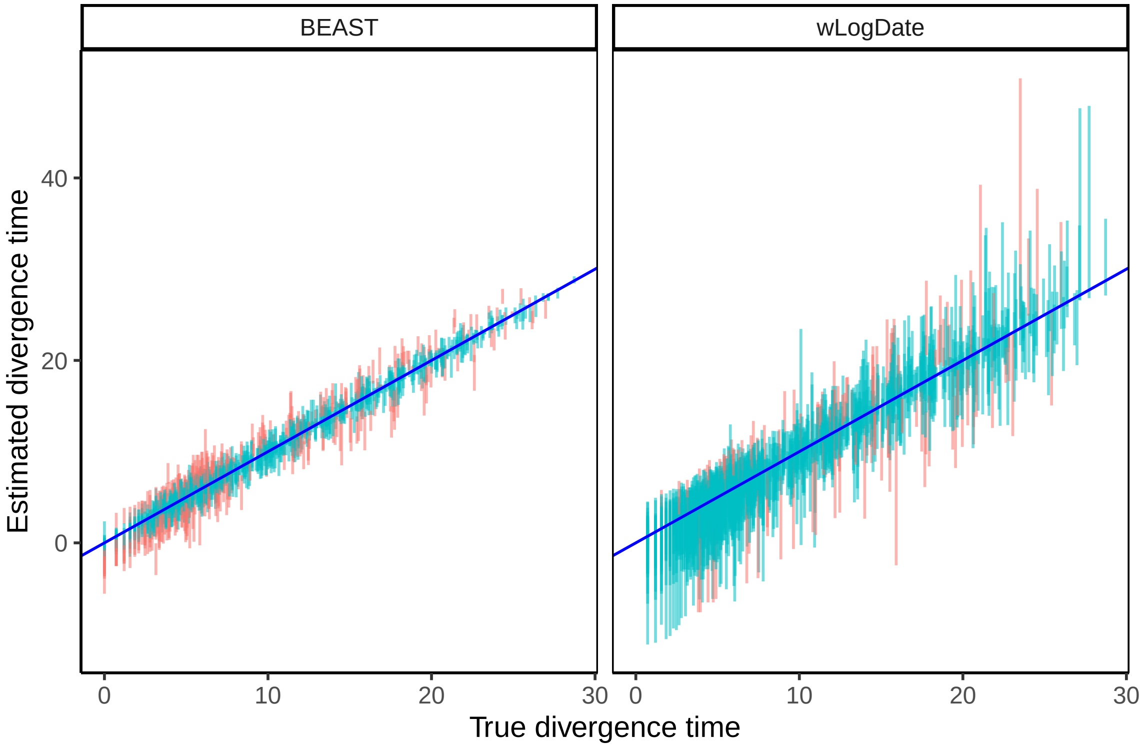


Figure 4.5: Estimated versus true divergence time. Each bar corresponds to the 95% confidence interval (CI) of one node estimate (each of the 109 nodes of the 10 replicates) by BEAST strict clock and wLogDate. Red color is used to mark points where the true time falls outside the CI.

varied estimates for the younger nodes (i.e. those with higher divergence times) and have more varied estimates for older nodes. In addition, the estimates of wLogDate are more concentrated around the true values than that of LF and RelTime, indicating a better accuracy. In two test cases (out of 200), LF had extremely high error (Fig. D.7 in Appendix D). Once those two cases are removed, the average RMSE normalized by tree height is 0.09 for wLogDate, 0.10 for LF, and 0.13 for RelTime (Table D7 in Appendix D). Comparing to LF and wLogDate, RelTime gives wider distributions of the estimates for a large portion of the nodes. Finally, the comparison in running time of wLogDate and RelTime is shown in Fig. D.8 (Appendix D).

Comparing to DAMBE in inferring unit time trees, wLogDate has lower error in $6/10$ replicates and DAMBE has lower error in the remaining 4 replicates (Table S8). Overall, the average error of wLogDate is 9.40%, which is slightly lower than that of DAMBE at 9.66%.

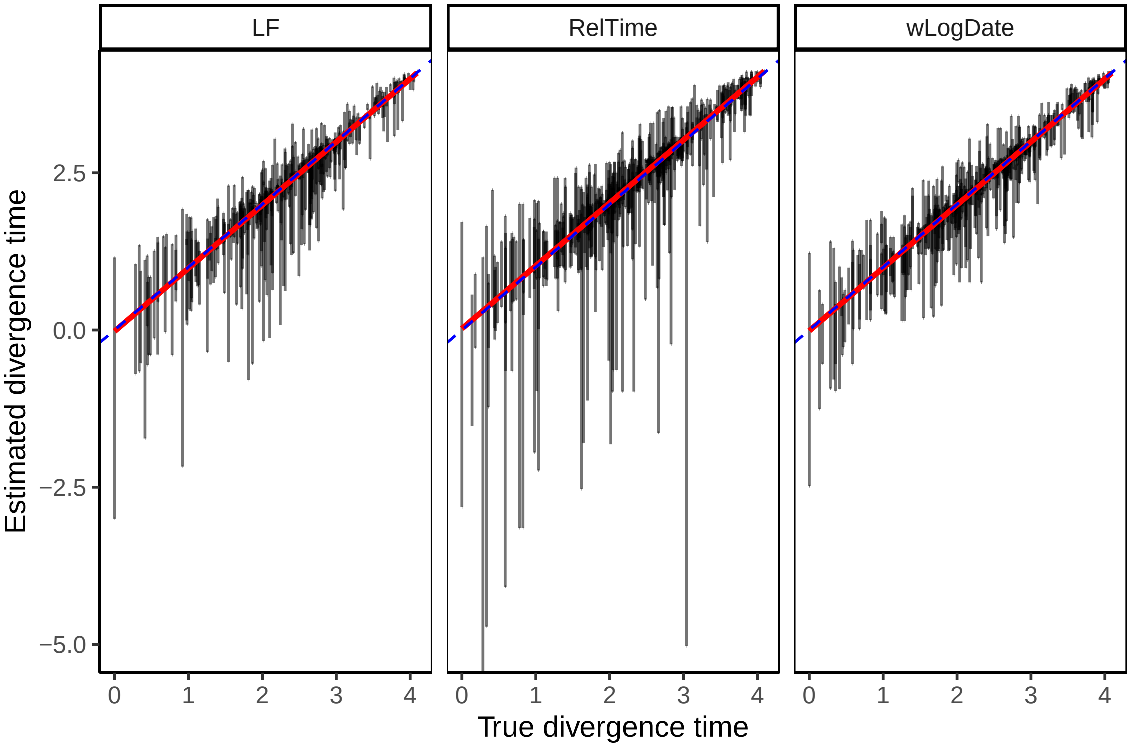


Figure 4.6: Comparison of LF, RelTime, and wLogDate on the simulated data with autocorrelated rate model. The y-axis shows estimated divergence times of uncalibrated internal nodes while the x-axis shows the true divergence time. Each bar shows the 2.5% and 97.5% quantiles of the estimates of a single node's divergence time across 20 tests, each of them with different random choices of calibration points (thus, these are not CIs for one run). There are 10 replicate trees, each with 44 uncalibrated nodes (thus, 440 bars in total). This figure discards 2 tests (out of $10 \times 20 = 200$) where LF produced extremely erroneous time trees (see Fig. D.9 in Appendix D for the full results). The root-mean-square error of the un-calibrated internal node ages, normalized by the tree height averaged across all replicates were 0.09, 0.1, and 0.13, respectively, for wLogDate, LF, and RelTime (see D.7 in Appendix D).

4.3.3 Biological data

On the H1N1 dataset, the best available evidence has suggested a tMRCA between December 2008 and January 2009 [184, 271, 127]. wLogDate inferred the tMRCA to be 14 December 2008 (Fig. 4.7a), which is consistent with the literature. LF and LSD both infer a slightly earlier tMRCA (10 November 2008), followed by BEAST-strict and BEAST-lognorm (October 2008 and July 2008), and finally BEAST runs using the phyML tree (Feb. 2008 for strict and July 2007 for LogNormal). While the exact tMRCA is not known on this real data, the results demonstrate that wLogDate, on a real data, produces times that match the presumed ground truth.

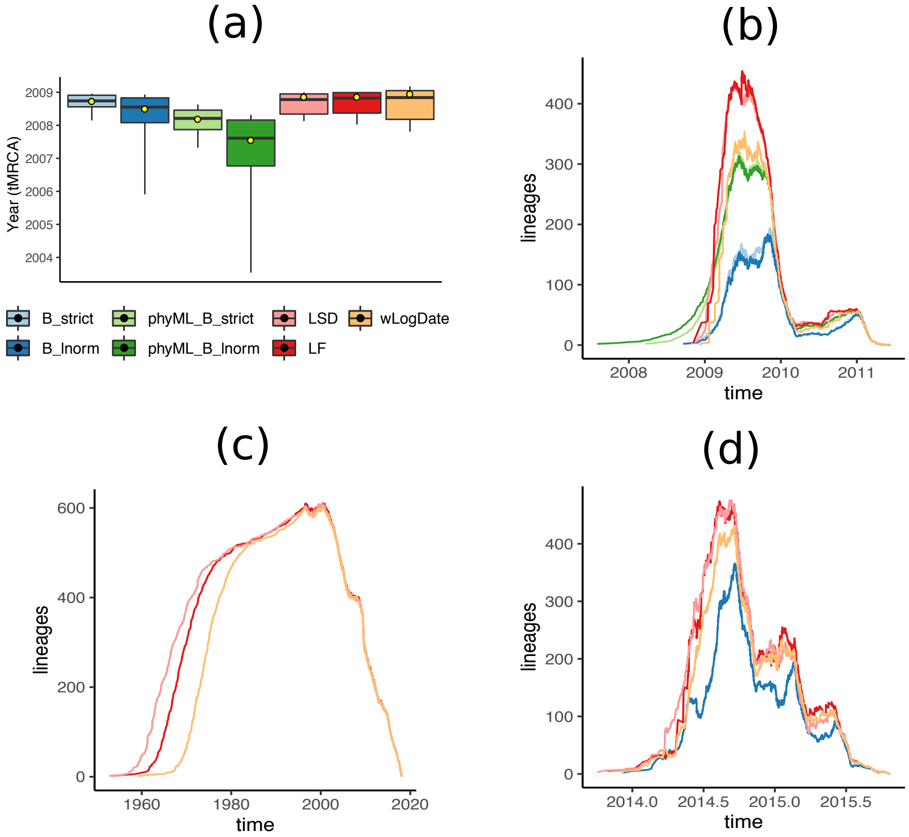


Figure 4.7: (a) Inferred tMRCA of the H1N1 dataset. Boxplots represent the median, maximum, minimum, 97.5% and 2.5% quantiles of the bootstrap estimates for LF, LSD, and wLogDate, and of the posterior distribution for BEAST. Yellow dot shows the inferred tMRCA of the best ML or MAP tree. BEAST was run with 4 different settings: B_strict and B_lnorm allow BEAST to infer both tree topology and branch lengths, with strict and LogNormal clock models; phyML_B_strict and phyML_B_lnorm fixed the topology to the rooted phyML tree given to BEAST. All other methods (LSD, LF, and wLogDate) were run on the rooted phyML trees. Results for LSD, LF, and BEAST are all obtained from [340]. (b) LTT plot for all methods on the H1N1 data. (c) LTT plot of fast methods on the HIV dataset. (d) LTT plot of BEAST, LSD, LF, and wLogDate on the Ebola dataset.

On the HIV dataset, wLogDate inferred a tMRCA of 1958 with only a handful of lineages coalescing in the 1950s and most others coalescing in 1960s and early 1970s (Fig. D.5 in Appendix D). The recovered tMRCAs is within the range postulated in the literature for subtype B [110, 350] and the fact that randomly sampled HIV lineages across USA tend to coalesce deep in the tree is a known phenomenon. LF and LSD recovered the tMRCA of 1952 and 1953, respectively.

Comparing to wLogDate, these two strict-clock methods postulate an earlier burst of subtype B (Fig. 4.7c). We were not able to run BEAST on this dataset.

On the Ebola dataset, the BEAST-1000 trees obtained from [80] inferred the tMRCA to be between 13 September 2013 and 26 January 2014 (95% credible interval) and the BEAST-MCC inferred the tMRCA to be 5 December 2013 as reported by [346]. Here, wLogDate inferred a tMRCA on 7 December 2013, which is very close to the estimate by BEAST. Both LF and LSD inferred an earlier tMRCA: 29 October 2013 for LF and 2 October 2013 for LSD, but still within the 95 per cent credible interval of BEAST-1000. LTT plots showed a similar reconstruction by all methods for this dataset (Fig. 4.7d).

We also compare running times of dating methods on the three real biological datasets (Table S3). LSD was always the fastest, running in just seconds, compared to minutes for LF and wLogDate. LF is faster than wLogDate on the H1N1 and HIV data, while on Ebola data, wLogDate is faster. We report the running time for wLogDate as the sequential run of 10 independent starting points and note that wLogDate can easily be parallelized. We further tested the scaling of wLogDate with respect to the number of species by subsampling the HIV dataset to smaller numbers of species (Fig. D.4 in Appendix D). The results show that the running time of wLogDate increases slightly worse than quadratically with the incrased number of species.

4.4 Discussion and future work

We introduced (w)LogDate, a new method for dating phylogenies based on a non-convex optimization problem. We showed that by log-transforming the rates before minimizing their variance, we obtain a method that performs much better than LSD, which is a similar method without the log transformation. In phylodynamics settings, our relatively simple method also outperformed other existing methods, including the Bayesian methods, which are much slower. The improvements were most pronounced in terms of the estimation of tMRCA and individual node ages and less so for the mutation rate. Moreover, improvements are most visible under the hardest model conditions, and are also observed in when data are generated according to autocorrelated model of rates.

The log transformation results in a non-convex optimization problem, which is harder to solve than the convex problems solved by LSD and LF. However, we note that the problem is convex for rate multipliers between 0 and e. In addition, given the advances in numerical methods for solving non-convex optimization problems, insistence on convex problems seems unnecessary. Our results indicate that this non-convex problem can be solved efficiently in the varied settings we tested. The main benefits of the log transformation is that it allow us to define a scoring function that assigns symmetrical penalties for increased or decreased rates (Fig. 4.1a); as we argued, this symmetry is a desirable property of the penalty function for several clock models that deviate from a strict clock.

The accuracy of LogDate under varied conditions we tested is remarkable, especially given its lack of reliance on a particular model of rate evolution. We emphasize that the parametric models used in practice are employed for mathematical convenience and not because of a strong biological reason to believe that they capture real variations in rates.

Even assuming biological realism of the rate model, the performance of the relaxed clock model used in BEAST was surprisingly low. For example, when rates are drawn from the LogNormal distribution, BEAST-strict often outperformed BEAST-LogNormal, especially in terms of the estimates of tMRCA and the mutation rate. We confirmed that the lower accuracy was not due to lack of convergence in the MCMC runs. We reran all experiments with longer chains (Table S5). to ensure ESS values are above 300 (Table S6). These much longer runs failed to improve the accuracy of the BEAST-LogNormal substantially and left the ranking of the methods unchanged (Fig. D.10).

The LogDate approach can be further improved in several aspects. First, the current formulation of LogDate assumes a rooted phylogenetic tree, whereas most inferred trees are unrooted. Rooting phylogenies is a non-trivial problem and can also be done based on principles of minimizing rate variation [201]. Similar to LSD, LogDate can be generalized to unrooted trees by rooting the tree on each branch, solving the optimization problem for each root, and choosing the root that minimizes the (w)LogDate objective function. We leave the careful study of such an approach to the future work.

Beyond rooting, the future work can explore the possibility of building a specialized solver for LogDate to gain speedup. One approach could be exploiting the special structure of the search space defined by the tree, which is the strategy employed by LSD to solve the least-squares optimization in linear time. Divide-and-conquer may also prove effective.

The weighting scheme used in LogDate is chosen heuristically to deal with the deviations of estimated branch lengths from the true branch length. In future, the weighting schema should be studied more carefully, both in terms of theoretical properties and empirical performance.

We described, implemented, and tested LogDate in the condition where calibrations are given as exact times (for any combinations of leaves and internal nodes). While the current settings fit well to phylodynamics data, its application to paleontological data with fossil calibrations is somewhat limited due to the requirements for exact time calibrations (in contrast to the ability to allow minimum or maximum constraints on the ages, or a prior about the distribution of the ages as in BEAST and RelTime). While the mathematical formulation extends easily, treatment of fossil calibrations and uncertainty of times is a complex topic [134, 125] that requires the expansion of the current work. Applying LogDate for deep phylogenies would need further tweaks to the algorithm, including changing equality to inequality constraints and the ability to setup feasible starting points for the solver.

In the studies of LogDate accuracy, we have explored various models for rate heterogeinety, including parametric models where rates are drawn i.i.d. from a fixed distribution (Log-normal, Exponential, and Gamma) and autocorrelated model where the rates of adjacent branches are correlated. Overall, none of the methods we studied is the best under all conditions.

In phylodynamics data, our simulations showed that it is more challenging for all the dating methods to date the phylogenies of the inter-host evolution (M3 and M4) than the intra-host (M1 and M2). wLogDate outperforms other methods for the inter-host phylogenies, regardless of the model of rate heterogeneity. While all methods have lower error for intra-host trees, BEAST with strict-clock prior is often the best method. However, the differences between BEAST and wLogDate are small and wLogDate is often the second best. Thus, wLogDate works well for virus phylogenies, especially in inter-host conditions. Despite the fact that RelTime explicitly optimizes the rate for each pairs of sister lineages, wLogDate is more accurate than both LF and RelTime on the data where the rates are autocorrelated between adjacent branches. These results show that wLogDate is applicable to a fairly large number of models of the trees and the rates.

Nevertheless, the approach taken by wLogDate suffers from its own limitations. By including a single mean rate around which (wide) variations are allowed, wLogDate is expected to work the best when rates have distribution that are close to being unimodal. However, rates on real phylogenies may have sudden changes leading to bimodal (or multimodal) rate distributions with wide gaps in between modes. For example, certain clades in the tree may have local clocks that are very different from other clades. Such a condition has been studied by [23] for a dataset of seed plants. The authors setup a simulation where there are local clocks on the tree and the mean values of these clocks are different by a factor varying from 3 to 6. [23] point out that under such condition, especially when the rate shift occurs near the root, BEAST usually overestimates the time of the Angiosperm (i.e. gives older time) by a factor of 2 (BEAST results from [23] are reproduced in Figure D.11 in Appendix D).

We also tested wLogDate, LF, and RelTime on this dataset (Fig. D.11 in Appendix D). In scenario 2 of the simulation, where the rate shift between the two local clocks is extreme (a factor of 6), wLogDate clearly over-estimate the age of Angiosperms (by a median of 55%). In this same scenario, RelTime slightly underestimate the age (by 5%). In the other 4 scenarios where the rate shifts are more gentle, wLogDate continue to overestimate the age but by small margins (by 6%, 1%, 2%, and 5%), while RelTime underestimates ages also by small margins (3%, 5%, 4%, 3%, and 3%). LF has similar patterns to wLogDate. These results point to a limitation of wLogDate (and the other dating methods) in phylogenies with multiple local clocks.

In addition to multiple clocks, future works should test LogDate under models where rate continuously change with time, and have a direction of change. Finally, to facilitate the comparison between different methods, we used the true topology with estimated branch lengths. Future work should also study the impact of the incorrect topology on LogDate and other dating methods.

Software availability The LogDate software is available on `https://github.com/uym2/wLogDate` in open-source format. The command-line python tool is available through conda for easy installation. A link to a web sever making wLogDate available as a web-server is also available from the github page.

Data availability All the data are available on `https://github.com/uym2/LogDate-paper`.

4.5 Acknowledgments

We thank Susan B. Little for providing the San Diego HIV sequence dataset used in this study.

This work was supported by the National Science Foundation (NSF) grant III-1845967 and a developmental grant from the University of California, San Diego Center for AIDS Research (P30 AI036214), supported by the National Institutes of Health.

Chapter 4, in full, is a reprint of the material as it appears in “Log Transformation Improves Dating of Phylogenies”. Mai, Uyen, and Siavash Mirarab. Molecular biology and evolution 38.3 (2021): 1151-1167. The dissertation author was the primary investigator and first author of this paper.

Chapter 5

MD-Cat: Phylogenetic dating under a flexible categorical model using Expectation-Maximization

One of the fundamental problems in phylogenetic reconstruction is dating, which is to convert branch lengths of a phylogenetic tree from the substitution unit to the time unit. While a small subset of node ages are known from either carbon-dating fossils or sampling times of the leaves, inferring the ages of the other nodes is nontrivial. Assuming a clock model (i.e. a model of how rates vary with time), we formulate dating as a maximum likelihood (ML) estimation problem. While there exists multiple ML-based methods addressing the same problem, their accuracy depends strongly on the type of the dataset, posing two challenges: (1) model misspecification: the assumed clock model that determines the likelihood function is misspecified and (2) non-convex optimization: the likelihood function involves an integral over continuous domain of the rates and is difficult to optimize. To tackle these two problems, we propose a new method called Molecular Dating using Categorical-models (MD-Cat) in this Chapter. MD-Cat is a nonparametric dating method using a categorical model of rates and the Expectation-

Maximization (EM) algorithm. We discretize the unknown rate distribution into k categories and approximate it by a categorical distribution. The EM algorithm is used to co-estimate the k rate categories and all the unknown branch lengths in time unit. Our model is free of any assumption about the true clock model, and has a sole parameter, k, that determines the resolution of the discretization. If k goes to infinity, this categorical model can in theory approximate any clock model as long as the unknown rates are i.i.d and there are infinite amounts of data available for estimation. On two simulated datasets of Angiosperms and HIV and a wide selection of rate distributions, MD-Cat is often more accurate than the alternatives, especially on datasets of either multiple local clocks or a global multimodal clock.

5.1 Introduction

Sequence data alone is not sufficient to infer phylogenetic branch lengths in the unit of time, so external sources of information are needed. Often, the external information comes in the form of *calibration points* such as ancestral divergence times that may be available from carbon-dating fossils or sampling times of tips available when studying phylodynamics. However, such external information alone is not sufficient either. Inferring a time tree requires assuming a *molecular clock* model of how the mutation rates have (or have not) changed across the tree.

The simple *strict clock* model [384], where the substitution rate is assumed to be constant across all tree branches, has been shown to be oversimplistic in some situations because mutation rates can vary substantially across tree branches [39, 170] especially at longer evolutionary time horizons. Attempts to relax the strict-clock have been made using both *uncorrelated* and *autocorrelated* models. With uncorrelated models, the rate on each branch is drawn independently from a common underlying parametric distribution such as exponential, gamma, or lognormal [15]. In autocorrelated models, the mutation rate evolves on the tree; i.e., the rate of each branch varies from that of its parent branch under a presumed model. There have also been efforts to model rate heterogeneity as sudden rate shifts throughout the evolutionary history, where a lineage possesses a dramatic rate change compared to its parent and passes on this new rate to its descendants, creating an entire subtree with a shifted rate. Such changes can happen multiple times on the tree creating local clocks with heterogeneous rates. Computational methods addressing these types of heterogeneous rates use a discrete clock model where a finite number of rate changes is allowed on the tree [14, 79, 101, 126, 373, 370]. With this assumption, these methods define local clock as a groups of taxa where every lineage evolves at exactly the same substitution rate. In addition, empirical studies have explored data with more complex and realistic heterogeneous clocks, such as a mixture of Lognormal distributions [23].

A large number of computational methods for molecular dating are available (see [171, 135, 283, 291]). Bayesian methods [273, 78, 119, 126, 339, 338], arguably the most popular set of methods, incorporate complex models in a Bayesian framework and use sampling methods to infer the time tree directly from sequence data. These methods enable the use of various types of priors for both molecular clock and calibration points, providing a way to make use of the prior knowledge. On the other hand, selecting the best priors can be difficult, and their merit has been recently debated [23, 351], especially when not given the correct clock prior. Moreover, the application of Bayesian methods can be hindered by the computational burden of the costly MCMC process.

An attractive alternative to Bayesian sampling is to assume a parametric model for the molecular clock and estimate the parameters of the model using maximum likelihood (ML) [177, 340, 346]. These ML-based methods usually do not work directly on sequence data, but take as input a tree in substitution units (such a tree can be estimated from sequence data using either distance-based [286, 105, 183] or ML-based methods [118, 320, 231]) and convert its branch lengths to time unit using a set of calibration points. Although they are often more efficient than the Bayesian methods and do not need a prior, ML-based methods have their own limitations. The simplest methods assume a strict clock model and use either a Poisson [177] or Gaussian [340] distribution to model the uncertainties of the (observed) branch lengths in substitution units. Compared to other methods, these strict clock methods are very fast and can handle phylogenies of millions of taxa. However, empirical evidence has now made it clear that mutation rates can vary substantially [39, 170] and therefore, these strict clock methods should only be used in limited contexts.

More sophisticated ML-based methods, such as TreeDater [346] and TreeTime [285], use a Gamma or Lognormal distribution to model the molecular clock. Because they assume a parametric – and often unimodal – clock model, these ML-based methods are not robust against model misspecification. The model violation is especially pronounced when the true clock model is multimodal – a phenomenon that can happen when the phylogeny has heterogeneous rates or multiple local clocks. In addition, the mutation rates in these models are often treated as continuous latent variables and the corresponding likelihood function involves an intractable integral over all possible rates in a continuous domain. Because such a likelihood function is difficult to optimize, some ML-based methods (e.g. [346]) have to depend on heuristic algorithms to iteratively optimize their likelihood functions. These heuristic algorithms lack theoretical supports, such as a guarantee to reach a local (or global) optimum, a monotonic improvement of the likelihood function, or a convergence guarantee.

There are also non-parametric methods for molecular dating [200, 290, 292, 331, 332, 367]. Most of these methods formulate dating as an optimization problem — often in a least-square form — to optimize a predefined objective function without an explicit parametric model of the clock. However, the objective functions used in these methods can be understood as making implicit assumptions about properties of the rate distribution. For example, many of these methods implicitly assume that rates are distributed following a unimodal distribution; that is, the rate of each branch is centering either around a global rate [367, 200] or the rate of its parent or sibling branch [290, 331, 37]. Incidentally, minimizing the residual sum-of-square often winds up being the ML solution under specific (often unimodal) models; e.g., our attempt at developing the non-parametric method wLogDate [200] produced a method that is the ML estimate under a set of (unimodal) LogNormal distributions. Thus, although most of the non-parametric methods are fast and show signs of robustness under multiple clock models, we postulate (and show in our study) that these non-parametric methods are inaccurate when the true clock model is multimodal.

We introduce MD-Cat, a new ML-based method for molecular dating. Unlike other ML-based methods, we use a categorical distribution (CAT model) to approximate the unknown continuous clock model. While categorical models have been adopted to approximate the Gamma model for rate heterogeneity across sequence sites [93, 320, 231], the power of the CAT model to approximate a continuous *unknown* clock model across branches has not been studied. The CAT model is non-parametric, in these sense that it does not assume the rates are drawn from a predefined parametric model. The model is defined as a set of k rate categories at unknown positions from which each rate is drawn. Although it is discrete, this CAT model has the power to approximate a continuous clock model if k is large and there are enough data to fit the model. We use the Expectation-Maximization (EM) algorithm to maximize the likelihood function associated with this model, where the k rate categories and the branch lengths in time units are treated as unknown parameters and are co-estimated. We show that both the E-step and M-step of our EM algorithm can be computed efficiently and the algorithm is guaranteed to converge. We then evaluate the method and compare it to the state-of-the-art on a wide range of simulated datasets.

5.2 Method

5.2.1 Notations

For a given binary tree T with n leaves and $N = 2n-2$ branches, we give each of the $n-1$ internal nodes of T a unique index in $\{0,\dots,n-2\}$ (reserving 0 for the root) and give each of the n leaf nodes a unique index in $\{n-1,\dots,N\}$. We denote the parent of node i as $par(i)$, the left and right children of node i as $c_l(i)$ and $c_r(i)$, and the edge connecting i and $par(i)$ as e_i for

$i \in \{1, \ldots, N\}$. The length of e_i is specified in either substitution unit or time unit. Let t_i denote the divergence time of node i (i.e. the time when species i diverges into $c_l(i)$ and $c_r(i)$). Then $\tau_i = t_i - t_{par(i)}$ is the length of e_i in time unit. As a shorthand, we combine all the τ_i values into a vector $\tau = [\tau_1 \tau_2 \ldots \tau_N]$. Similarly, let b_i be the length of e_i in substitution unit, which is the *expected number of substitutions per sequence site* occurred on edge e_i and let $\mathbf{b} = [b_1 b_2 ... b_N]$. We assume that the mutation rate can change only at species divergence times and let $\mu_i = b_i / \tau_i$ denote the rate along branch e_i. Finally, we let s denote the alignment length.

5.2.2 The generative model

We assume sequences of length s evolve from the ancestral sequence at the root of a binary tree T with mutation rate μ_i along each edge e_i. We assume μ_is are identically and independently drawn (i.i.d) from an unknown distribution and sites of the molecular sequences evolve i.i.d along each edge e_i following a homogeneous process such as the GTR model [336]. Using methods such as Maximum Likelihood (ML), we can obtain an estimate of the topology of T and all branch lengths $\mathbf{b} = [b_1 b_2 ... b_N]$ in substitution unit.

The Gaussian model of branch length estimation error

We model the uncertainties in branch length estimation using a Gaussian model. Let B_i be a random variable denoting the estimated length of branch e_i in substitution unit and $\hat{b_i}$ be the *observed* estimated branch. The distribution of B_i depends on several factors such as the sequence evolution model, sequence length, and the inference technique. For simplicity, we model B_i using the Gaussian model as in [340]: let $\varepsilon_i = B_i - b_i$ be the estimation noise, we assume $\varepsilon_i \sim_{i.i.d.} N(0, \frac{b_i}{s})$, where the variance b_i / s comes from an approximation of the Poisson model of the number of substitutions per sequence site (see [340] for details). In addition, for computational and algorithmic convenience, we approximate the variance $\frac{b_i}{s})$ by $\frac{\hat{b_i}}{s}$ and add a pseudo-count of $1/s$ to $\hat{b_i}$ to account for zero-event branches (see [340] and [200] for discussions about this issue). Recall that $b_i = \mu_i\tau_i$ and $B_i = \varepsilon_i + b_i$, so we have $B_i \sim_{i.i.d.} N(\mu_i\tau_i, \frac{\hat{b}_i}{s})$. Thus,

$$f(\hat{b}_i|\mu_i,\tau_i) = \frac{1}{\sqrt{2\pi\hat{b}_i/s}} \exp\left(-\frac{1}{2\hat{b}_i/s}(\hat{b}_i - \mu_i\tau_i)^2\right) . \tag{5.1}$$

Approximation of the unknown clock model

As mentioned previously, the clock model that describes the distribution of μ is uncertain. Common choices include Gaussian [340, 285], Gamma [346], LogNormal and Exponential distributions [77]. However, there is no guarantee that these distributions, which happen to be unimodal or mode-less, would be good models of how μ changes across a tree. For example, a bimodal or tri-modal set of rates would not be adequately modeled with any of these models.

Instead of using a parametric continuous distribution, we use a non-parametric approach using a discrete distribution. We discretize μ into k categories $\omega = [\omega_1, \omega_2, \dots \omega_k]$ each with the same probability mass $\frac{1}{k}$ and assume μ_is are i.i.d. under this k-categorical model. We use this categorical model to approximate the unknown distribution that μ was drawn from.

5.2.3 Maximum likelihood estimation using EM algorithm

Under the model described in sections 5.2.2 and 5.2.2, τ and ω are parameters, $\hat{b}_i$'s are observations (i.e. data), and μ_i's are latent variables. Following the ML inference framework, we find the unknown parameters τ and ω that maximize the log-likelihood function and satisfy a set of constraints defined by calibration or sampling times. We then employ the EM algorithm to solve this optimization problem.

The linear constraints defined by calibration points

Let t_0 be the *unknown* divergence time at the root of the tree. The p calibration points $t_1, \ldots, t_p$ define a set of p constraints $C_1, ..., C_p$:

$$C_i : \sum_{j \in P(0,i)} \tau_j = t_i - t_0 \tag{5.2}$$

where $P(x,y)$ denote the path between two nodes x and y (referring to by their indices). Recall that the root has index 0, so $P(0,i)$ is the path from the root to node i. To remove t_0 from this set of constraints, we arbitrarily select a constraint C_k and subtract it from the other constraints C_i ($i \neq k$) side by side to obtain the set $\Psi^{(k)}$ of $p-1$ linear constraints on τ:

$$\Psi_i^{(k)} : \sum_{j \in P(\mathrm{LCA}(i,k),i)} \tau_j - \sum_{j \in P(\mathrm{LCA}(i,k),k)} \tau_j = t_i - t_k, \tag{5.3}$$

It is easy to see that all the linear constraint sets $\Psi^{(k)}$ ($k \in [p]$) are equivalent. Therefore, we use Ψ as a short hand to refer to any (arbitrarily chosen) linear constraint set among them. We can construct a set Ψ using a bottom up traversal of the tree, as in [200].

The log-likelihood function

Under the categorical model of the rates and the Gaussian model of branch length estimation error, the log-likelihood of $\hat{b}_i$'s given the parameters τ and ω is

$$\begin{aligned} l(\tau, \omega) &= \textstyle\sum_{i=1}^{N} \log L_i(\hat{b}_i; \tau_i, \omega) \\ &= \textstyle\sum_{i=1}^{N} \log \sum_{j=1}^{k} f(\hat{b}_i; \omega_j, \tau_i) Pr(\mu_i = \omega_j; \omega) \\ &= \textstyle\sum_{i=1}^{N} \log \left(\frac{1}{k} \sum_{j=1}^{k} f(\hat{b}_i; \omega_j, \tau_i) \right) \end{aligned} \tag{5.4}$$

where $L_i(\hat{b}_i; \tau_i, \omega)$ denotes the density of $\hat{b}_i$ on branch i and f is the density function of the Gaussian model described in section 5.2.2. Our goal is to find τ and ω that maximize $l(\tau, \omega)$ and satisfy Ψ. Because l has a summation inside the log function, it is difficult to directly optimize l. However, thanks to the categorical model of the rates, the latent variables μ_i are discrete and so we can readily apply the EM algorithm [69] to maximize l, as shown below.

EM-based optimization

In the EM algorithm [69], we start with an initial of τ and ω (described later) and iteratively improve the log-likelihood function by alternating between the E-step and M-step.

E-step In the E-step, we compute the posterior of the latent variables:

$$q_{ij} = Pr(\mu_i = \omega_j | \hat{b}_i; \tau, \omega) = \frac{f(\hat{b}_i|\mu_i=\omega_j,\tau_i)Pr(\mu_i=\omega_j;\tau,\omega)}{\sum_{m=1}^{k} f(\hat{b}_i|\mu_i=\omega_m,\tau_i)Pr(\mu_i=\omega_m;\tau,\omega)} = \frac{f(\hat{b}_i|\mu_i=\omega_j,\tau_i)}{\sum_{m=1}^{k} f(\hat{b}_i|\mu_i=\omega_m,\tau_i)}$$

The second equality holds because $Pr(\mu_i = \omega_m; \tau, \omega) = \frac{1}{k}$ for all i, m.

M-step In the M-step, we find ω and τ to maximize

$$\sum_{i=1}^{N} \sum_{j=1}^{k} q_{ij} \log f(\hat{b}_i|\omega_j, \tau_i), \tag{5.5}$$

s.t. the constraints Ψ are satisfied (see Eq. (5.3)). Using Eq. (5.1) and removing constants, we can reduce the problem to:

$$\min_{\tau,\omega} \sum_{i=1}^{N} \sum_{j=1}^{k} \frac{sq_{ij}}{\hat{b}_i} (\hat{b}_i - \omega_j \tau_i)^2 \tag{5.6}$$

s.t. $\omega > 0$, $\tau > 0$, and Ψ are satisfied.

Solving the M-step The optimization problem in the M-step (equivalently defined by Eqs. (5.5) and (5.6)) is non-convex and is difficult to solve exactly. However, it is easy to see that the likelihood function is bounded above, so it has a maximum. Therefore, the EM algorithm still converges as long as after every iteration (h) the M-step finds a new point $(\tau^{(h+1)}, \omega^{(h+1)})$

that gives a higher value for Eq. (5.5) (or equivalently, lower value for Eq. (5.6)) than that of $(\tau^{(h)}, \omega^{(h)})$. In other words, it is sufficient to find a local minimum of Eq. (5.6) in every iteration and guarantee the convergence of the EM algorithm (a proof of convergence is shown in Supplementary).

We use coordinate descent to find multiple local minima of Eq. (5.6) and select the one that gives the lowest penalty. In the M-step, starting with $\omega^{(h)}$ and $\tau^{(h)}$, we successively minimize Eq. (5.6) along the coordinate block of either ω or τ while fixing the other, and iterate until convergence. In other words, let $\tau^{(h,1)} = \tau^{(h)}$ and $\omega^{(h,1)} = \omega^{(h)}$; in each iteration (p) of coordinate descent, we compute $\tau^{(h,p+1)}$ and $\omega^{(h,p+1)}$ such that:

$$\tau^{(h,p+1)} = \underset{\tau \geq 0, \tau \text{satisfies} \Psi}{\arg\min} \sum_{i=1}^{N} \sum_{j=1}^{k} \frac{sq_{ij}}{\hat{b}_i} (\hat{b}_i - \omega_j^{(h,p)} \tau_i)^2 \tag{5.7}$$

$$\omega^{(h,p+1)} = \underset{\omega \geq \varepsilon}{\arg\min} \sum_{i=1}^{N} \sum_{j=1}^{k} \frac{sq_{ij}}{\hat{b}_i} (\hat{b}_i - \omega_j \tau_i^{(h,p+1)})^2 \tag{5.8}$$

The above two problems are instances of the *weighted least-square* optimization; they are convex and can be solved efficiently using any canonical convex programming algorithm. Inspired by the algorithms presented in [340], we use the active-set method to solve these two problems, and we show that the complexity of each iteration of the active-set method is $O(k)$ (Appendix E).

Initialization The EM algorithm can only find local optima, so it needs multiple initial points. To facilitate the search, we first estimate the expected mutation rate μ, then we discretize the uniform distribution $[0, 2\mu]$ into k equal segments and set $\omega_j^{(0)}$ to the middle of the j^{th} segment. In other words,

$$\omega_j^{(0)} = (2j-1)\frac{\mu}{k}, \forall j \in [k] \tag{5.9}$$

To initialize $\tau_i^{(0)}$, we draw j uniformly in $[k]$ and set $\tau_i^{(0)} = \frac{\hat{b}_i}{\omega_j^{(0)}}$. Although in this initialization $\tau_i^{(0)}$ does not satisfy Ψ, the constraints will be satisfied after the first M-step. We run wLogDate [200] and root-to-tip (RTT) regression [304] independently to get two different estimates of μ. We then use each of these two estimates of μ to get m different initials for $\tau^{(0)}$, for a total of $2m$ initials. In all experiments in this paper, unless otherwise specified, m is set to 100 and so EM is run with $2m = 200$ initials.

The 2-round optimization strategy A focal difficulty in the co-estimation of rates and times is that they are inseparable as a product. This same problem occurs in the EM algorithm, where in each iteration ω and τ can both be scaled up or down by the same factor. To avoid a rapid jump of ω and τ, we first enforce ω to have an expected value of the initial μ (estimated by either wLogDate or RTT, as described above) and run EM until convergence. Then we relax that constraint and let EM continue running (to allow the expected μ to be re-estimated). In other words, for each initial point, we run EM twice:

- In the first round, the expected mutation rate is fixed to the initial μ by adding the constraint $\sum_j \omega_j = k\mu$ to Ψ. The EM algorithm is run until convergence.
- In the second round, we initialize EM with the ω and τ found in the first round and relax the constraint $\sum_j \omega_j = k\mu$. We let EM algorithm run until convergence.

The final solution is the one that gives the lowest value for Eq. (5.6) among all initial points.

5.3 Simulated data

5.3.1 Angiosperms hybrid rate

Beaulieu *et al.* [23] simulated a hybrid rate model for a phylogeny of seed plants in which evolutionary rates formed local clocks in certain clades of the tree. The authors simulated 5 different scenarios where they change the relative ratios between some clades in the tree, as follow:

- scenario 1 = 3:1 herbaceous to woody angiosperm
- scenario 2 = 6:1 herbaceous to woody angiosperm
- scenario 3 = 4:1 angiosperm to gymnosperm; 3:1 herbaceous to woody angiosperm
- scenario 4 = 4:1 angiosperm to gymnosperm; 3:1 herbaceous to woody angiosperm; Gnetales herbaceous angiosperm
- scenario 5 = 4:1 angiosperm to gymnosperm; 3:1 herbaceous to woody angiosperm; Gnetales woody angiosperm

The time tree and 100 simulated phylograms for each of these five scenarios were downloaded from the Dryad Repository provided by the authors. We used the provided phylograms to simulate DNA sequences of length 1000 using SeqGen under GTR model with the shape of the Gamma rate heterogeneity across sites set to $\alpha = 1$.

We compare MD-Cat to two other methods: wLogDate [200] and RelTime [331]. We keep the true rooted tree topology fixed and only infer branch lengths. We use RAxML [320] to estimate the branch lengths in substitution unit from the simulated sequences (using the GTRGAMMA model) and use MD-Cat, wLogdate, and RelTime to infer the time tree. Because these dating methods require rooted tree as input, we use the 20 species on the clade outside the Angiosperm as outgroups to root the RAxML tree. As outgroup rooting cannot determine the exact root position on the branch connecting ingroups and outgroups, we remove this entire clade from the tree. As such, the 5 calibration points that belong to the 20 species in the outgroups are discarded, leaving us with 15 calibration points to be used in this analysis. This same setting is used for RelTime, wLogDate, and MD-Cat. We provide each method with the exact time of those 15 calibration points. We run wLogDate with 100 feasible starting points, MD-Cat with $k = 50$ rate categories and 200 initial points, and RelTime with default settings.

5.3.2 HIV phylodynamics

We reuse the phylodynamics data of HIV *env* gene simulated in previous analyses [340, 200], but we explore many more clock models. The time trees were simulated based on a birth-death model with periodic sampling times. There are four tree models: D995_11_10 (M1) and D995_3_25 (M2) simulate intra-host HIV evolution and D750_11_10 (M3) and D750_3_25 (M4) simulate inter-host evolution. Each tree model has 100 replicates, so we have 400 different tree topologies in total. Using these time trees, in our earlier work [200], we simulated phylograms (i.e. the phylogenetic trees with branch lengths measured in substitution unit) using three clock models: Lognormal, Gamma, and Exponential. In this paper, we further augment this dataset by nine new clock models, one of which is a uniform distribution and the other eight are mixtures of two, three, or four Lognormal distributions. We also simulate sequence data using Seqgen under the same settings as the original study [340] and our prior work [200]: sequence length is 1000; DNA evolution model is the F84 model with a gamma distribution for across-sites rate heterogeneity with shape 1.0 and eight rate categories; transition/transversion rate ratio is 2.5; nucleotide frequency of (A, C, G, T) is $(0.35, 0.20, 0.20, 0.25)$. Table 5.1 summarizes the parameters and statistics of all 12 clock models used in this paper.

We compare MD-Cat to wLogDate [200] and BEAST [79]. In this study, we keep the true rooted tree topology fixed and only infer branch lengths in time unit. To test wLogDate and MD-Cat, we use RAxML [320] to estimate the branch lengths in substitution unit from the simulated sequences (using the GTRGAMMA model) and use each of these methods to infer the time tree. wLogDate was run with 100 feasible starting points. MD-Cat was run with $k = 50$ rate categories and 200 initials. To test BEAST, we use the sequences simulated by SeqGen, also fix the true rooted tree topology and only inferred node ages. We run BEAST using the following priors: HKY+Γ8 model, coalescent with constant population size, and strict-clock (i.e. fixed rate) clock model. We set the length of the MCMC chain to 10^7 generations, burn-in to 10%, and sampling to every 10^4 generations.

Model name	Parameters	Mean	Std	CV	Newly simulated
LogNormal	$\mu = 0.006, \sigma = 0.0024$	0.006	0.0024	0.4	No
Gamma	$\alpha = 6.05, \beta = \alpha/\mu$ where $\mu = 0.006$	0.006	0.00244	0.407	No
Exponential	$\lambda = 1/\mu$ where $\mu = 0.006$	0.006	0.006	1.0	No
Uniform	$a = 0, b = 0.012$	0.006	0.0035	0.577	Yes
Bimodal 1	$\mu_1 = 0.003, \sigma_1 = 0.0003, p_1 = 0.5$ $\mu_2 = 0.009, \sigma_2 = 0.0003, p_2 = 0.5$	0.006	0.003	0.5	Yes
Bimodal 2	$\mu_1 = 0.002, \sigma_1 = 0.0003, p_1 = 0.5$ $\mu_2 = 0.01, \sigma_2 = 0.0003, p_2 = 0.5$	0.006	0.004	0.667	Yes
Bimodal 3	$\mu_1 = 0.003, \sigma_1 = 0.0024, p_1 = 0.5$ $\mu_2 = 0.009, \sigma_2 = 0.0024, p_2 = 0.5$	0.006	0.0038	0.641	Yes
Bimodal 4	$\mu_1 = 0.002, \sigma_1 = 0.0024, p_1 = 0.5$ $\mu_2 = 0.01, \sigma_2 = 0.0024, p_2 = 0.5$	0.006	0.0047	0.783	Yes
Trimodal 1	$\mu_1 = 0.002, \sigma_1 = 0.0003, p_1 = 0.2$ $\mu_2 = 0.006, \sigma_2 = 0.0003, p_2 = 0.6$ $\mu_3 = 0.01, \sigma_3 = 0.0003, p_3 = 0.2$	0.006	0.00254	0.423	Yes
Trimodal 2	$\mu_1 = 0.002, \sigma_1 = 0.0003, p_1 = 0.4$ $\mu_2 = 0.006, \sigma_2 = 0.0003, p_2 = 0.2$ $\mu_3 = 0.01, \sigma_3 = 0.0003, p_3 = 0.4$	0.006	0.0036	0.6	Yes
Trimodal 3	$\mu_1 = 0.002, \sigma_1 = 0.0003, p_1 = 0.333$ $\mu_2 = 0.006, \sigma_2 = 0.0003, p_2 = 0.333$ $\mu_3 = 0.01, \sigma_3 = 0.0003, p_3 = 0.333$	0.006	0.003	0.5	Yes
Quartmodal	$\mu_1 = 0.001, \sigma_1 = 0.0003, p_1 = 0.25$ $\mu_2 = 0.004, \sigma_2 = 0.0003, p_2 = 0.25$ $\mu_3 = 0.008, \sigma_3 = 0.0003, p_3 = 0.25$ $\mu_4 = 0.011, \sigma_4 = 0.0003, p_4 = 0.25$	0.006	0.0038	0.633	Yes

Table 5.1: Parameters and statistics of the 12 clock models. Lognormal distributions are parameterized by μ and σ, which are the *actual* mean and standard deviation of the distribution. The bimodal, trimodal, and quartmodal distributions are mixtures of $2,3$, or 4 Lognormal distributions, respectively, and p_i is the probability mass of component i of the mixture. Gamma distribution is parameterized by its shape α and rate β. The other distributions are shown by their canonical parameterization.

5.4 Results

5.4.1 Angiosperms hybrid rate

We compare MD-Cat, wLogDate, and RelTime on their accuracy in estimating the age of the common ancestor of Angiosperms. In this simulation by [23], the true age in all replicates is fixed to 140 mya. We show the distribution of the estimates across 100 replicates for each

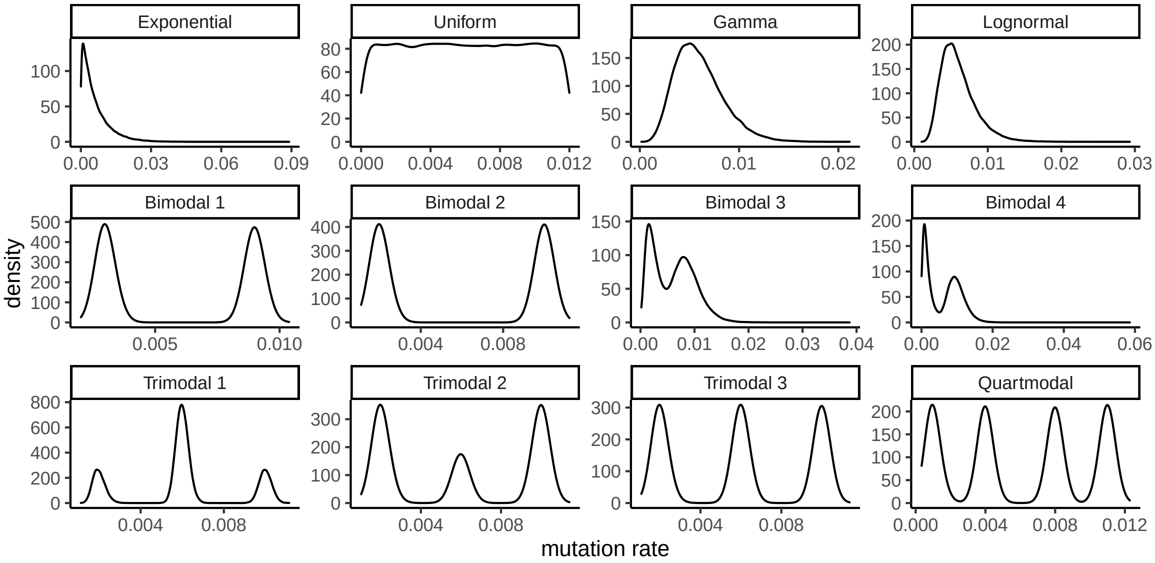


Figure 5.1: The empirical distribution of the mutation rates simulated for each of the 12 clock models used in the HIV dataset.

methods in five scenarios (Fig. 5.2a). wLogDate tends to overestimate the Angiosperms age for all scenarios, by a median of 12, 88, 15, 13, and 13 million years for scenarios 1–5, respectively, and has very high error (35% in average). Compared to wLogDate, RelTime has lower error (6.2% in average); however, it significantly underestimates the age in all 5 scenarios, by a median of 8, 10, 10, 8, and 9 million years for scenarios 1–5, respectively. MD-Cat, in contrast, has both low error (5.2% in average) and low bias: for scenario 1, the method overestimates the age by 5 million years; for scenarios 2–5, it underestimates by 0.5, 1.6, 0.1, and 0.6 million years, respectively. Comparing the estimated branch lengths in time unit (Fig. 5.2b), wLogDate seriously underestimates the shorter branches, but tends to overestimate the longer branches, especially in scenario 2. RelTime has an overall tendency to underestimate the branch lengths (consistent with the underestimation of tMRCA). MD-Cat overestimates the shortest branch in all 5 scenarios and slightly overestimates the short branches in scenario 1, but overall MD-Cat has tighter distribution of the estimates compared to wLogDate and RelTime, and is less biased.

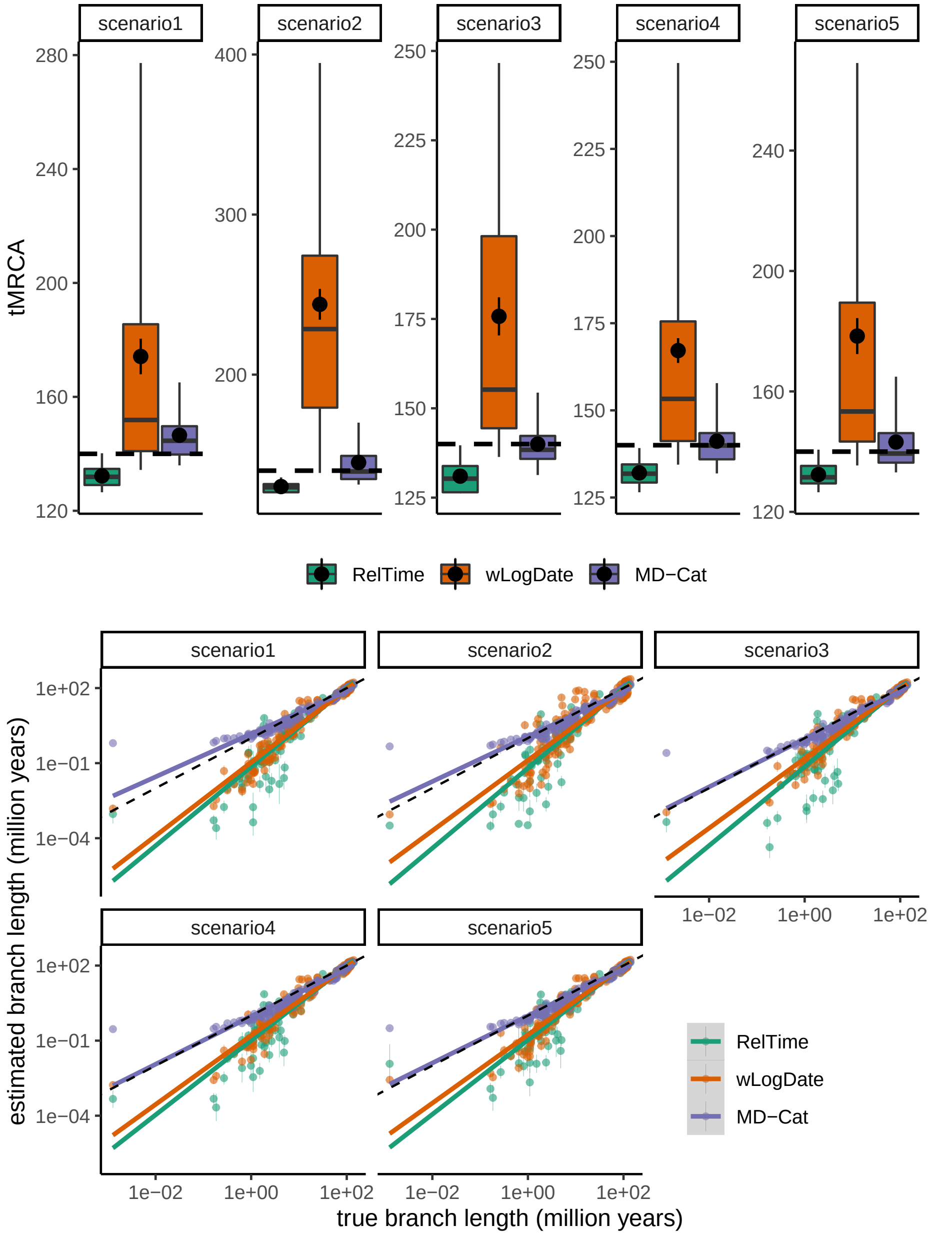


Figure 5.2: Comparison of MD-Cat, wLogDate, and RelTime on the Angiosperms simulated dataset. Top: the tMRCA estimated by each method on each of the 5 scenarios. The dashed-line shows the true tMRCA at 140. Bottom: the estimated versus true branch lengths in time unit. Mean and standard error are shown for the estimates across 100 replicates per branch. Both axes are shown in log-scale. Dashed line shows the unity line (i.e., perfect accuracy).

5.4.2 HIV phylodynamics

First, we compare MD-Cat, wLogDate, and BEAST (with strict-clock prior) on their accuracy in estimating the divergence time of the root (Fig. 5.3 Top). Both wLogDate and BEAST overestimate the tMRCA in all clock models. The average bias of wLogDate is 4.8% and that of BEAST is 9.8% across all models. MD-Cat overestimates the tMRCA in 7 models (Exponential - 6.7%, Gamma - 3.4%, Lognormal - 3.0%, Bimodal 3 - 3.5%, Trimodal 1 - 2.0%, Trimodal 3 - 1.6%, and Uniform - 2.9%) and slightly underestimates the other 5 (Bimodal 1 - 0.6%, Bimodal 2 - 1.4%, Bimodal 4 - 0.07%, Trimodal 2 - 0.5%, and Quartmodal - 0.4%). Averaging across all models, MD-Cat has a small overestimation bias of 1.7%. Next, we compare the normalized error (i.e. Root-Mean-Squared-Error divided by the tree height) of all the estimated divergence times (Fig. 5.3 Bottom). In the two unimodal models (Gamma and Lognormal), BEAST and wLogDate have similar accuracy and are both more accurate than MD-Cat, albeit by a small margin (for the Lognormal model, the average error by BEAST, wLogDate, and MD-Cat are 4.2%, 4.1%, and 4.9%, respectively; for the Gamma model, the average error by BEAST, wLogDate, and MD-Cat are 4.4%, 4.2%, and 5.0%, respectively).

In contrast, MD-Cat is more accurate than the other two methods in the other clock models. Across all clock models, the average error of wLogDate, BEAST, and MD-Cat are 8.0%, 8.6%, and 6.7%, respectively. Importantly, MD-Cat gives a tighter error distribution than that of BEAST and wLogDate (both for the MRCA and all other nodes divergence times), demonstrating its robustness across all replicates and clock models. At the extremes, BEAST and wLogDate can have an error of 40% or more in some replicates (e.g. in Exponential and bimodal 4), while MD-Cat maintains an overall low to moderate 95-percentile error that is below 20% across all clock models (Fig. 5.3 Bottom).

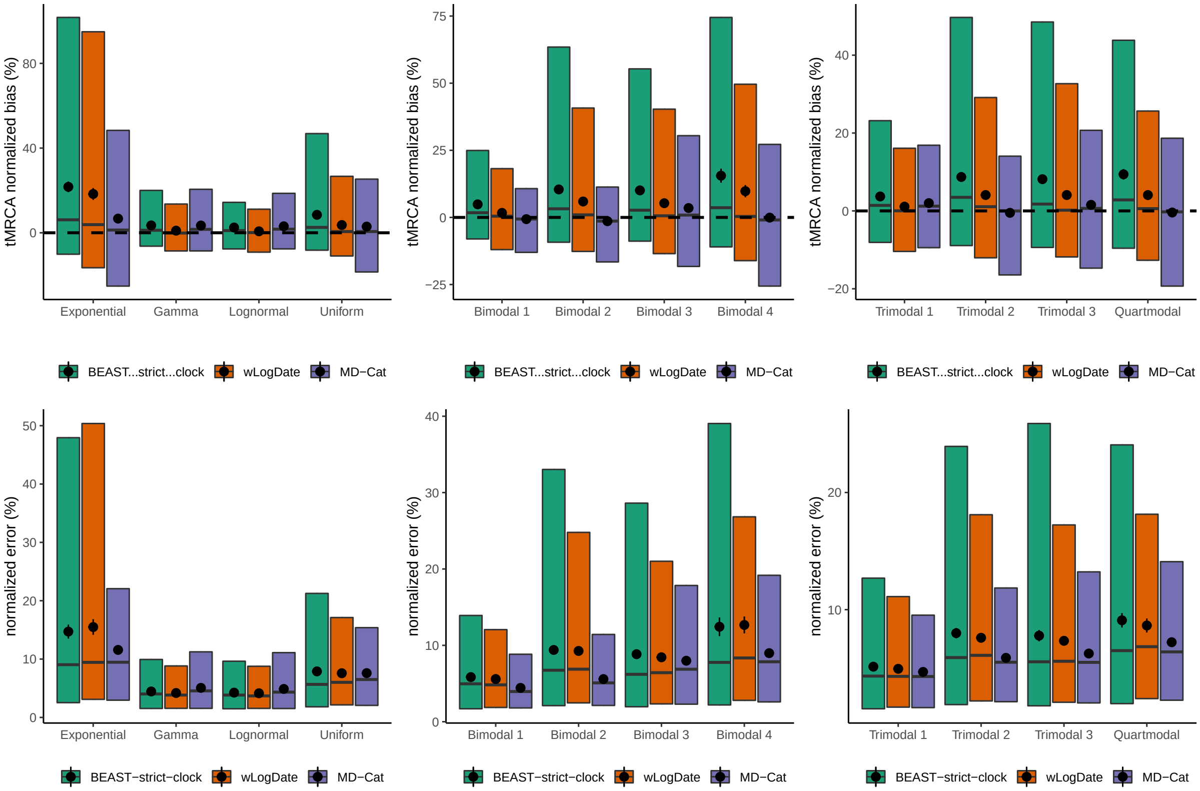


Figure 5.3: Comparison of MD-Cat, wLogDate, and BEAST on the HIV simulated dataset. Top: the bias in estimated root divergence time (tMRCA). Bottom: normalized error (RMSE divided by tree height) of estimated divergence times for all nodes. All boxplots show 5-95 percentiles across all 400 replicates. Horizontal line shows median. Mean and standard error are also shown as a dot and error bars. Model conditions are divided into those with zero or one mode (left), two modes (middle), and three or more modes (right).

5.5 Discussions

We introduced MD-Cat, a new method for dating phylogenies using a categorical model. Although the categorical model is not new in phylogenetics, its power to approximate a continuous distribution as the clock model has not been studied before. We formulate the dating problem under this cateogrical model in a maximum likelihood estimation framework and show that the problem can be solved effectively using an EM-based algorithm. Although the likelihood function is non-convex, the objective function in the M-step is convex on ω and τ separately. We used coordinate descent to alternatively optimize ω and τ, and show that each iteration can be solved in linear time. Together, these strategies greatly reduce the running time of the MD-Cat method.

On two simulated datasets of the Angiosperms and HIV, the MD-Cat method outperformed other methods in most clock models, especially in the model conditions that violate their (implicit) assumptions of a strict or unimodal clock model. Improvements are most visible under the hardest model conditions (i.e. scenario 2 in the Angiosperms dataset where the rate shift is extreme and the Exponential and bimodal 4 in the HIV dataset where the rate distribution has high variance). Compared to other methods, MD-Cat is more robust; it consistently maintains the 95-percentile error under 20%.

The formulation of MD-Cat can face the problem of over-parameterization. As noted before, time and rate are generally inseparable from molecular data, even with the aiding information from calibration points. In general, approximating a continuous distribution requires a large number of rate categories (k). However, the data size is very limited compared to the number of unknown parameters ($2n-2$ branch length observations and $O(n)$ calibration points versus $2n-2+k$ parameters). Thus, these parameters are unidentifiable in general, pointing to the difficulty of getting the correct age for every node. With the pre-estimate of the expected mutation rate and our 2-round optimization strategy (described in section 5.2.3), we have tried to address the over-parameterization and improved the accuracy of MD-Cat. Thus, our evidence that the inference of a substantial number of rates is doable despite the large space is mostly empirical rather than theoretical. We leave the analysis of identifiability of ω and τ under this categorical model and different combinations of calibration points for future work.

While we formulate and estimate ω and τ under a maximum likelihood framework, we note that the proposed categorical model can also be used in Bayesian inference, where a prior for the rate categories can be specified. In addition, for scalability purposes, we use the inferred tree in substitution unit as input and use a simple Gaussian model for branch length estimation error. However, a more direct method is to incorporate the categorical clock model into the tree inference from sequence data and maximize the joint likelihood function. We note that this approach avoids the assumption of the Gaussian error in branch length estimation used in our study, but at the same time it may exacerbate the over-parameterization problem, as all the parameters (tree topology, GTR parameters, per-site and per-branch rate heterogeneity) must all be co-estimated in a maximum likelihood framework. Future works should explore this approach, both theoretically and empirically.

The current formulation of MD-Cat assumes a rooted phylogenetic tree. However, we note that rooting phylogenies is a non-trivial problem and also related to the problem of clock model selection. A straight-forward generalization of MD-Cat to unrooted trees is to solve the optimization problem for each possible rooting and select the root position that has the maximum likelihood. Such an approach should be explored in future work, together with an updated formulation for MD-Cat to add a parameter that determines the optimal root placement on each branch.

In addition to ways to improve the method, our study can also be extended. To facilitate the comparison between different methods, we used the true topology with estimated branch lengths and left the study of the impact of incorrect topology on the dating methods for future works. Finally, testing under more complex clock models, such as those that allow continuously changing rates is also worth further examining.

Chapter 6

Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea

Rapid growth of genome data provides opportunities for updating microbial evolutionary relationships, but this is challenged by the discordant evolution of individual genes. Here we build a reference phylogeny of 10,575 evenly-sampled bacterial and archaeal genomes, based on a comprehensive set of 381 markers, using multiple strategies. Our trees indicate remarkably closer evolutionary proximity between Archaea and Bacteria than previous estimates that were limited to fewer “core” genes, such as the ribosomal proteins. The robustness of the results was tested with respect to several variables, including taxon and site sampling, amino acid substitution heterogeneity and saturation, non-vertical evolution, and the impact of exclusion of candidate phyla radiation (CPR) taxa. Our results provide an updated view of domain-level relationships.

6.1 Introduction

The metaphor of a "tree of life" was used by Darwin in his *On the Origin of Species* in 1859. It came into its modern form when Carl Woese and co-workers used the new ability to genetically sequence the small subunit (SSU) ribosomal RNA gene from multiple different organisms to create a phylogenetic tree [359], thereby showing a scenario of three domains of life: Bacteria, Archaea, and Eukaryota [360, 243]. Recent years have seen discoveries of novel microbial groups enabled by culture-based and metagenomic methods [276, 233, 359, 383], many of which represent previously unknown biodiversity [40, 377, 276], and keep updating our knowledge of the extent and relationships among domains as indicated by phylogenetics [58, 144, 121, 50]. Among these new discoveries is the candidate phyla radiation (CPR, also referred to as Patescibacteria) [276, 40], a highly diversified clade of mainly uncultivated microorganisms that may subdivide the domain of Bacteria [144], although this scenario remains controversial [246]. Meanwhile, the discovery and analysis of multiple novel archaeal lineages have suggested an archaeal origin for eukaryotes, pointing to a two-domain scenario [356, 86]. The currently representative view of the tree of life, inferred based on the concatenation of ribosomal proteins, illustrated a bipartite pattern with distinct separation between Bacteria (including CPR) and Archaea (plus Eukaryota) [144, 50]. More comprehensive work in both taxon and locus sampling exists [246], but the inter-domain relationships were not explored.

Reconstructing phylogenies typically relies on comparing homologous features. Although closely related organisms often share obvious genome-level homologies, building higher-level, especially cross-domain phylogenies has been challenging due to the rarity of clearly defined homologies [239]. To date, many efforts rely on one, or a few, universal "core" genes that are usually involved in fundamental translation machinery [356, 124]. Examples include the SSU rRNA [54, 264, 111], and several dozens ribosomal proteins [272]. The choice of these "marker genes" is based on their universality, conservation, and the observation that they suffer from less frequent horizontal gene transfer (HGT) [239]. However, HGT is widespread across the domains [54, 264, 111], affecting even the most conserved “housekeeping” genes [59], and cannot be ruled out even for these markers. Furthermore, the reliance on a few marker genes limits the information (i.e., phylogenetic signal) available for resolving all relationships in the tree of life. Finally, it reduces applicability in metagenomics—increasingly the main source of novel genome data—where assembled genomes are frequently incomplete and error-prone [359]. Maximizing the included number of loci, thus, is desirable. However, when dealing with many loci, reconciling discordant evolutionary histories among different parts of the genome can become challenging. Moreover, a practical dilemma is imposed by computational limitations: adding breadth across the phylogenetic space requires more computing effort, which leads to compromises with either the quantity of genes analyzed [144] or the robustness of tree-building algorithms [246].

In this work, we build a reference phylogeny of 10,575 bacterial and archaeal genomes (Fig. 6.1). They are sampled from all 86,200 non-redundant genomes available from NCBI GenBank and RefSeq [122] as of March 7, 2017 (Fig. 6.2), using a statistical approach that maximizes the covered biodiversity. Our phylogenetic reconstruction uses 381 marker genes, selected from whole genomes solely by sufficient sequence conservation to identify homology. The whole data set totals 1.16 trillion non-gap amino acids, making it among the largest single data sets upon which *de novo* phylogenetic trees have been built (Supplementary Table F.1). To infer species trees, we use both a summary approach that accounts for discrepancy among the evolutionary histories of individual genes, and the conventional gene alignment concatenation approach. The resulting species trees provide high resolution of the basal relationships among microbial clades, which show that Bacteria and Archaea are in closer proximity compared to previous estimations (Fig. 6.1). The phylogeny also enable us to evaluate and revise previously established taxonomic hierarchies. We have made our data and protocols publicly available at https://biocore.github.io/wol/.

6.2 Results

6.2.1 Comprehensive sampling of biodiversity and genes

By using a purpose-built “prototype selection” algorithm to maximize evenness of genome sampling (Supplementary Fig. F.1, detailed in Supplementary Note F.1.1) and by incorporating multiple additional criteria, including marker gene presence, genome quality, and taxonomy, we selected 10,575 genomes, covering 146 of 153 phyla defined by NCBI, plus all 89 classes, 196 of 199 orders, 422 of 429 families, 2,081 of 2,117 genera, and 9,105 of 20,779 species (Fig. 6.2a). A total of 2,852 genomes (27.0%) are metagenome-assembled genomes (MAGs) while the remaining are from isolates and other sources (Fig. 6.2c). Meanwhile, 2,267 genomes (21.4%) are complete genomes or chromosomes, while the remaining are scaffolds or contigs (Fig. 6.2d). Overall, the selected genomes are of high completeness and low contamination as evaluated based on known lineage-specific marker gene sets (Fig. 6.2b). By testing against the MAG quality standard established by Bowers *et al.* [359], only 10.4% MAGs or 3.7% of all genomes fall within the low-quality draft category, while the remaining meet the criteria of either high or medium quality drafts (Fig. 6.2e). This balanced representation of known bacterial and archaeal diversity ensured the comprehensiveness and evenness of the resulting phylogeny.

Our phylogenomic analysis was based on the 400 marker genes originally proposed in PhyloPhlAn [303] (Supplementary Fig. F.2). The taxon sampling protocol ensured that all selected genomes contain at least 100 marker genes each. In the resulting data matrix, each marker gene is present in 7,565 ± 1,730 (mean and std. dev.) genomes (Supplementary Fig. F.2a), while each genome contains 286.14 ± 80.23 (mean and std. dev.) marker genes (Supplementary Fig. F.2b). These marker genes were further filtered down to 381, based on metrics of alignment quality (see Methods) across the sampled genomes (Supplementary Fig. F.2d).

6.2.2 Assessing deep phylogeny using multiple strategies

We explored multiple tree inference methods (detailed in Supplementary Note F.1.2, with selected ones compared in Fig. 6.3 and Supplementary Fig. F.3), but will mostly focus on two strategies: CONCAT and ASTRAL. CONCAT concatenates gene alignments and infers a single tree using maximum likelihood (ML) performed using the robust implementation in RAxML [321]. Computational limitations forced us (Supplementary Table F.2) to use at most 100 sites per gene, selected either randomly ("concat.rand") or based on maximum conservation ("concat.cons"). However, we also tested analyzing all sites, using the faster but less accurate ML program, FastTree [261] (referred to as "fasttree"). In contrast, the ASTRAL tree ("astral") is based on first inferring 381 gene trees and then summarizing them using the ASTRAL method [378]. ASTRAL accounts for gene tree discordance due to divergent coalescent histories and has been shown in simulations to be more accurate than concatenation in the presence of highly frequent HGTs [66]. Due to its inherent scalability, ASTRAL analyses were able to use all the data (i.e., all sites of every gene). For comparison to previous studies [144], a CONCAT tree was also built using 30 ribosomal proteins ("concat.rpls"). We used ML to estimate branch lengths for the ASTRAL tree based on the same data used to infer the CONCAT tree.

Overall, ASTRAL (Fig. 6.1, Supplementary Fig. F.4) and CONCAT trees (Supplementary Figs. F.5 and F.6) show congruence in topology (Fig. 6.3a, b, Supplementary Note F.1.2) when compared to trees derived from implicit (e.g., distance-based) analyses (Supplementary Fig. F.7, Supplementary Note F.1.2). The congruence is higher at shallow branches, but generally decreases as phylogenetic depth increases (Supplementary Fig. F.8). The ASTRAL tree, in particular, has high support among the early branching clades (Supplementary Figs. F.3 and F.9, also see Supplementary Figs. F.4-F.6). This high resolution is directly related to the large number of gene trees used in the inference, as using fewer loci notably decreased the branch support of the species tree (Supplementary Fig. F.10, Supplementary Note F.1.2). On the other hand, the evolutionary relationships recovered by CONCAT are impacted by the breadth of site sampling (Supplementary Fig. F.11, Supplementary Note F.1.2) and the robustness of method (Supplementary Fig. F.12, Supplementary Note F.1.2).

To further evaluate the impact of taxon sampling, we tested a series of subsampled sets of taxa, selected so that they maximize the representation of large and deep-branching clades (see Methods). Reducing taxon sampling changed the overall topology (Supplementary Fig. F.13, Supplementary Note F.1.2) and the inferred relationship between large groups (e.g., placement of Chloroflexi and Chlamydiae) (Supplementary Fig. F.14), further highlighting the importance of our dense sampling of genomes.

Phylogenetic trees built by both strategies recapitulated clear separation between Archaea (669 taxa) and Bacteria (9,906 taxa) at the root (Figs. 6.1 and 6.3, Supplementary Figs. F.4-F.6). Meanwhile, CPR (1,454 taxa) forms a monophyletic group located at the base of the bacterial lineage in the ASTRAL tree and the CONCAT trees that use global markers (Figs. 6.3c and 6.4a). Considering the potential impact of long branch attraction, this placement will require further validation using more robust substitution models and controlled tests. The ASTRAL tree shows high consistency and moderate-to-high branch support for several taxonomic units recently defined above the phylum level, including TACK, Microgenomates, Parcubacteria, FCB, PVC, and Terrabacteria [276] (Fig. 6.3c, d). These groups were also supported in the CONCAT trees, with the exception of Terrabacteria in one analysis (Fig. 6.3c, d). With reference to the trees, we systematically evaluated and curated NCBI taxonomy, showing frequent incongruences (Supplementary Fig. F.15a, c, Supplementary Table F.3), especially in metagenome-derived genomes (detailed in Supplementary Note F.1.3). We further compared our trees with the recently developed GTDB taxonomy and trees [246], and observed overall high congruence, though with a few exceptions at deep branches (Fig. 6.3a-d, Supplementary Figs. F.15b, d, F.16, elaborated in Supplementary Note F.1.4). A detailed interpretation of our phylogeny in reference to taxonomy and multiple previous works is provided in Supplementary Note F.1.5.

6.2.3 Evolutionary proximity between Archaea and Bacteria

ASTRAL and CONCAT trees both reveal a relatively short branch connecting the most recent common ancestors of Archaea and Bacteria (Figs. 6.1 and 6.4a, c, Supplementary Fig. F.17). Its length is fractional comparing to the dimensions of both clades (appr. 0.13-0.14 by conserved sites, 0.09-0.11 by random sites) (Fig. 6.4c, e, Supplementary Table F.4). This pattern is in contrast to previous trees built using fewer marker genes, all or most of which are ribosomal proteins formerly considered to be effective markers for assessing global microbial evolution [272] (e.g., [50, 54, 376]). To further test how the choice of marker genes affects the inter-domain distance, we estimated branch lengths of the ASTRAL tree using 30 ribosomal proteins extracted from the genomes. Consistent with previous studies, we observed an elongated branch connecting Bacteria and Archaea. Its length relative to clade dimensions (1.0-1.6) is about 10 times the estimate using the 381 global marker genes (Fig. 6.4b, d and e, Supplementary Table F.4). We also calculated the overall phylogenetic distance between taxa of the two domains, as relative to the intra-domain distances. This relative distance based on the ribosomal proteins (4.5-5.0) is around three times that of the distance by the global marker genes (1.5-1.6) (Fig. 6.4f, Supplementary Table F.4).

Considering the special status of CPR, we performed an independent test with the 1,454 CPR genomes removed from the data set prior to *de novo* phylogenetic inference, and we compared the results to the main results (Fig. 6.4e, f) with the CPR clade pruned from the tree. These trees continued to reveal the substantially shorter branch and tip-to-tip distances between the two domains as recovered by using the 381 global marker genes as compared to using the 30 ribosomal proteins (Supplementary Fig. F.18, Supplementary Table F.5).

We tested whether the potential saturation of amino acid substitution could cause an underestimation of the domain separation. The ratio between phylogenetic distance and sequence distance is similar between pairs of taxa selected both from Bacteria, both from Archaea, or one from each domain (Supplementary Fig. F.19). This indicates that the relative length of the branch connecting the two domains compared to the intra-domain branches is not substantially impacted by saturation.

We further evaluated how individual gene trees impact the observed proximity between Bacteria and Archaea. Except for a few outliers, which include several "core" genes like *rpoC* (RNA polymerase subunit β', 18.27), *tuf* (elongation factor Tu, 12.18) and *fusA1* (elongation factor G, 9.54), most gene trees have relative Archaea-Bacteria distances between 1 and 3 (mean: 2.00) (Fig. 6.5a, b), which is consistent with that of the species tree summarizing the global marker genes, and in contrast to that obtained using only the ribosomal proteins (Fig. 6.5a).

6.2.4 Heterogeneity among individual genes' evolutionary histories

Because microbial genomes are highly dynamic and prone to HGTs, it is important to investigate the discrepancies among the evolutionary paths of individual gene families to better understand the evolution of genomes [264]. To measure the topological concordance between two trees, we used the quartet score [297], which correlates well with the traditional Robinson–Foulds (RF) metric (Supplementary Fig. F.20c) [279], resulting in a distribution of gene trees tightly centered around the species tree (Fig. 6.5d, Supplementary Figs. F.20a and F.21).

The discordance between the 381 single gene trees and the species tree varied widely (Fig. 6.5c). The quartet scores (larger is more similar, with identical trees scoring 1.0) ranged from 0.372 (cmpD) to 0.943 (hslU), with the mean and standard deviation being 0.653 ± 0.136. Many of the individual trees with high similarity to the species tree belong to genes involved in the core machinery of genetic information processing, such as those encoding DNA/RNA polymerase subunits, ribosomal proteins, and elongation factors, while genes involved in peripheral functions such as membrane transport are frequently more distant from the species tree (Fig. 6.5c, d). This pattern is generally consistent with a previous study on a small taxon set [264]. While determining the cause of discordance for individual genes is beyond the scope of this study, the pattern we observed is consistent with a reduced rate of HGT for fundamental genes compared to those with less conserved functional significance [148]. There was no apparent correlation between a gene tree's concordance with the species tree and the prevalence of the gene in the sampled genomes (Supplementary Fig. F.20d, e), suggesting that universality is not necessarily indicative of fidelity.

To further test the impact of gene tree discordance on the species tree, we sequentially removed genes from the low end of the quartet score rank (Supplementary Note F.1.2). ASTRAL produced stable topologies in this test (Supplementary Fig. F.22a-c). We next tested the impact on phylogenetic distances. There was a weak positive correlation (Pearson correlation R2 = 0.157, p = 1.88e-07) between the quartet score and the relative Archaea-Bacteria distance (Fig. 6.5e). When the branch lengths of the species tree were estimated using genes with high quartet scores only, the distance moderately increased, yet remained far from the result by using the ribosomal proteins (Supplementary Table F.6). This suggests that non-vertical transmission of genetic information has only a limited impact on our updated estimates of the inter-domain distance.

6.2.5 Heterogeneity across sites

Inferring phylogenetic trees at deep time scales, beyond the heterogeneity of gene histories, requires paying attention to the heterogeneity of substitution processes across the genome [255, 179]. As recently as 2015, Gouy *et al.* declared the jury to still be out on the root of the tree of life [115] , partially due to difficulties in modeling heterogeneity of sequence evolution across sites. In particular, changes in amino acid frequency across sites of the same gene can exacerbate long branch attraction [178]. To account for these difficulties, we tested whether our main conclusions stand if the data are analyzed with a recently developed model, PMSF, which considers heterogeneity in the amino acid substitution process [348] (Fig. 6.3: "pmsf.cons" and "pmsf.rand"). Because of the computational complexity of this approach, we had to limit these analyses to 1,000 taxa. At this sampling depth, we were also able to build a tree using all sites and the CONCAT method for comparison ("concat.al1k").

The topology of PMSF trees largely resembled the RAxML trees with the same taxon sampling (Fig. 6.3a, b). The impact of using the PMSF model instead of site homogeneous models on the topology and branch lengths was small compared to the impact of taxon, locus, and site sampling (Supplementary Fig. F.23, Supplementary Note F.1.2). The PMSF trees continued to support a large portion of relationships among deep branches recovered by the full-scale trees (Fig. 6.3c, d, Supplementary Fig. F.3). The evolutionary proximity between Bacteria and Archaea continued to hold with the PMSF trees. Meanwhile, the PMSF tree based on ribosomal proteins ("pmsf.rpls") also resembled the corresponding full-scale tree in suggesting a long distance between Bacteria and Archaea (Fig. 6.4e, f, Supplementary Fig. F.24, Supplementary Table F.7). Taken together, this shows that our phylogenies and main conclusions were robust when considering site heterogeneity.

6.3 Discussion

The origin and evolution of life has been among the most intriguing scientific questions, with the current widely adopted notion being the three-domain system: Bacteria, Archaea and Eukaryota [360]. Recent phylogenomics studies typically indicated a long distance between Bacteria and Archaea, with Eukaryota as an ingroup of the Archaea clade [58, 144]. In this work, we built a reference phylogeny of over 10,000 bacterial and archaeal genomes, covering a significant proportion of the known biodiversity with available genome data. The result provides an updated view of microbial evolution, showing that Bacteria and Archaea, the two microbial domains conventionally but controversially grouped by the term "prokaryotes" [354], are much closer in evolutionary proximity than estimates using a smaller number of "core" genes, such as the ribosomal proteins. This observation was further supported by extensive analyses using multiple tree-building methods, with consideration of taxon and site sampling, amino acid substitution heterogeneity and saturation, and non-vertical evolution, and was robust against the exclusion of

CPR taxa. Interestingly, applying a simple universal molecular clock as well as relaxed clock rates to date our trees resulted in divergence time estimates of major lineages that are compatible with geological timeline only when using the global markers, but not when trees are restricted to ribosomal proteins (Fig. 6.6, Supplementary Figs. F.25 and F.26, Supplementary Tables F.8-F.10, see full details in Supplementary Note F.1.6). These comparisons suggest accelerated evolution in ribosomal proteins during the separation between Bacteria and Archaea. They show the limitation of using core genes alone to model the evolution of the entire genome, and highlight the value in using a more diverse marker gene set.

Our work highlights the value of even taxon sampling, a global marker gene set representing the larger average of genome content, and comparative phylogenetic analyses. These procedures largely reduced the bias of gene choice in exploring genome evolution, and allowed us to characterize the evolutionary discrepancies of individual gene families. Despite these efforts, some lineages are still underrepresented in our sampling, such as DPANN [276] , which has genomes that are often missing many of the 381 marker genes (detailed in Supplementary Note F.1.5). Moreover, the rapid growth of genomic data has led to the absence of some newly discovered groups from our tree. While it is impractical to repeat all of our analyses to include all new genomes, it is of interest to assess whether the newly discovered microbial diversity may impact our results. Prior to submission of this article, we updated the genome collection from NCBI on May 23, 2019, and selected 187 new genomes representing phyla as defined by the newest NCBI and GTDB taxonomies that are absent or underrepresented in the current set of 10,575 genomes (see Methods). Phylogenetic trees built using the extended genome set continued to support the domain-level relationships in both topology and evolutionary distance as recovered by the main analysis (Supplementary Fig. F.27, Supplementary Table F.11, see Supplementary Note F.1.7 for full details). Finally, we note that the inclusion of eukaryotes is challenging with the current marker gene set due to the overall sparsity of detectable homology. Further improvements in methodology are important in order to deliver a robust phylogeny that encompasses all forms of life.

6.4 Methods

6.4.1 High-performance computing

Analyses of the genome datasets in this study were computationally intensive. Heavy computations used the “Comet” supercomputer located at the San Diego Supercomputer Center (SDSC). Each standard node is equipped with 24 Intel Haswell CPU cores and 128 GB of DDR3 memory, while each GPU node is equipped with four NVIDIA P100 GPUs, plus 28 Broadwell CPU cores and 128 GB memory. A small proportion of the computations used the “Barnacle” computer system operated by the Knight Lab, each node of which has 32 Haswell CPU cores and 256 GB DDR3 memory. Whenever possible, all CPU cores were used in a typical multi-processing task to minimize run time. Tasks that required more than 128 or 256 GB memory used the large-memory nodes of Comet, featuring 64 CPU cores and 1.5 TB memory per node. Benchmarking of the prototype selection algorithms and some local developments were performed on the “WarpDrv” workstation, equipped with 32 Intel Sandy Bridge CPU cores and 256 GB of DDR3 memory.

6.4.2 Retrieval of genome data and metadata

Microbial genomes were downloaded from the NCBI genome database (GenBank and RefSeq) as of March 7, 2017. We used and provided updates related to this work to the automated workflow RepoPhlAn (https://bitbucket.org/nsegata/repophlan, commit 03f614c) to download genomes from the NCBI server. Each genome was given a unique identifier, which was derived from the NCBI accession of the corresponding assembly but without version number. For example, a genome with assembly accession “GCF_000123456.1” was identified as “G000123456” in this study. In cases when the same genome was present in both GenBank (accession starting with "GCA_") and RefSeq (accession starting with "GCF_"), only the RefSeq version was kept.

6.4.3 Annotation and classification of marker genes

The functional annotation of the 400 PhyloPhlAn marker genes [303] was performed by aligning the protein sequences of the 400 marker genes (inferred from 2,887 bacterial and archeal genomes as described in Segata *et al.* [303]) against the UniRef50 database (March 2018 release) using BLASTp. The best hit for each gene was taken and queried against the UniProt database for gene and protein names. To categorize genes by function, the UniProt entries were translated into Gene Ontology (GO) terms [17] with the "subset_prokaryote" tag (March 2018 release). Because not all UniProt entries have corresponding GO terms, manual curation was involved to pick the most appropriate GO terms for those cases by examining the BLAST hit table. GO terms were further translated into GO slim terms to obtain higher-level functional categories. Note that this analysis is independent from the phylogenetic analysis of the current genome data set, and the result can be used as a reference for PhyloPhlAn users.

6.4.4 Analyses of genome sequences and identification of marker genes

The DNA sequences of the 86,200 bacterial and archaeal genomes were subjected to the following analyses:

1 The quality scores for DNA, protein, tRNA and rRNA were calculated following Land *et al.* [176].

2 A MinHash sketch was built for each genome using Mash 1.1.1 [240], with default settings (sketch size = 1000, *k*-mer size=21), based on which a pairwise distance matrix was built for the entire genome pool. In brief, MinHash is a k-mer hashing technique that enables the

quantification of genome-to-genome distance. It is efficient for very large genome sets, and it correlates well with the conventionally used average nucleotide identity (ANI) [240].

3 Although NCBI provides genome annotations, we chose to re-annotate the genomes using a uniform protocol to ensure consistency. Specifically, open reading frames (ORFs) were predicted using Prodigal 2.6.3 [147], in the single-genome mode, and allowing ORFs to run off edges of scaffolds.

4 Based on the predicted ORFs, the 400 marker genes were inferred and extracted using the phylogenomics pipeline PhyloPhlAn (commit 2c0e61a) [303], in which the 400 marker genes were originally introduced. In brief, we used USEARCH v9.1.13 to align ORFs against the reference marker gene sequences (see above) at an E-value threshold of 1e-40. It then selected the highest bit score hit of each ORF. Should more than one hit per marker per genome was observed, the highest bit score hit was selected as the representative of that marker gene.

5 The completeness, contamination, and strain heterogeneity scores were computed using CheckM 1.0.7 [247] with the default protocol ("lineage_wf").

6.4.5 Prototype selection and genome sampling

Proper taxon sampling is a key prerequisite to inferring an unbiased view of organism evolution [226, 81]. Beyond computational challenges in robust tree-building, the highly uneven distribution of known biodiversity (e.g., 40.0% of all genomes (34,507) belong to the nine most-sequenced species) requires deliberate subsampling to reduce the bias from the resulting phylogeny in representing a global view of evolution. We therefore applied the data-reduction strategy of "prototype selection" [104], which subsamples genomes from the pool such that they represent the largest possible biodiversity—in terms of maximized sum of pairwise distances as defined by k-mer signatures (Supplementary Fig. F.1a). We developed a heuristic (detailed in

Supplementary Note F.1.1), capable of handling the size of the current genome pool, with results comparable to or better than published alternatives (Supplementary Fig. F.1b-e).

Using this algorithm and by applying multiple criteria, we downsampled the 86,200 bacterial and archaeal genomes to 11,079. The procedures are detailed below.

1 Excluded genomes with marker gene count < 100 or contamination $> 10\%$. The marker gene count threshold 100 was chosen because it is sufficiently large to yield high resolution of the tree using ASTRAL (Supplementary Fig. F.10a, c). The contamination threshold 10% is inline with the medium- and low-quality draft genome standards proposed by Bowers *et al.* [359]. Nevertheless, we did not adopt the completeness and tRNA/rRNA coverage thresholds [359], because the 400 protein-coding marker genes are more relevant for phylogenetic reconstruction.

2 Included the NCBI-defined reference and representative genomes (https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/).

3 Included genomes that are the only representative of each taxonomic group from phylum to genus.

4 Included genomes that are the only representative of each species without defined lineage (no classification other than species).

5 Executed the prototype selection algorithm developed in this work: "destructive_maxdist" (see Supplementary Note F.1.1) based on the distance matrix defined by MinHash signatures, with the already included genomes as seeds, to obtain a total of 11,000 genomes.

6 For each phylum to genus, and species without classification from phylum to genus, selected one with the highest marker gene count. This added 79 genomes to the selection.

These 11,079 genomes were subjected to our phylogenetics protocol, during which further filtering was performed based on sequence alignment quality (see below). Eventually, 10,575 genomes were retained.

6.4.6 Impact of alternative genetic codes

We chose to uniformly apply the standard archaeal and bacterial genetic code table 11 to all genomes in order to minimize bias. Reports have shown that several lineages, such as Mycoplasma/Spiroplasma [158], Hodgkinia [208] and Absconditabacteria [46], use alternative genetic code tables 4, 25 and others, in most of which a stop codon is repurposed to encode for an amino acid, resulting in ORF elongation. We did not incorporate alternative genetic codes, however, because there is no accurate way to associate each of the 86,200 genomes with its true genetic code. Incorrect truncation of ORFs may unnecessarily exclude genes and taxa, whereas incorrect elongation of ORFs could result in artificially long branches because the amino acid sequence after a true stop codon is likely relaxed from selective pressure. Considering our goal of inferring phylogenetic topology and distances, we decided to only use the standard genetic code.

However, we did test the impact of using alternative genetic code on the gene and taxon sampling. We ran Prodigal 3.0.0-rc1, which automatically switches from genetic code 11 to 4 if the average ORF length is too short. This resulted in altered gene calling results in 453 out of the 86,200 genomes, of which 63 had overly short ORF lengths even when using genetic code 4. PhyloPhlAn marker gene discovery on the other 390 genomes with genetic code 4 suggested marginal increase in the extracted number of the 400 marker genes per genome (1.23 ± 5.28, mean and std. dev.). Only seven additional genomes which had less than 100 marker genes managed to pass this threshold (see above) after switching to genetic code 4. Therefore, omitting alternative genetic code has little impact on the inclusion of genomes.

6.4.7 Metric multidimensional scaling (mMDS) of genome distances

The effect of prototype selection was visualized using the mMDS technique, which renders a low-dimensional plot that minimizes the loss of information when transforming from the high-dimensional data. We performed mMDS using the “mds” function implemented in scikit-learn 0.19.2 [250] on the genome distance matrix, using the default setting, to compute the coordinates at the top five axes. The resulting coordinates were visualized with the interactive tool Emperor [345] as bundled in QIIME 2 release 2017.12 [30].

6.4.8 Protein sequence alignment and filtering

Protein sequences of each of the 400 marker gene families were aligned using UPP v2.0 [227], a phylogeny-based and fragmentary-aware alignment tool. UPP consists of several sequentially connected modules. It first identifies suspected fragmentary sequences, then calls PASTA v1.8.0 [214] to align the remaining sequences and build a phylogeny (backbone tree) based on them. Then it builds an ensemble of HMMs using HMMER [82] based on the phylogeny. Finally, it aligns the fragmentary sequences to the HMMs and selects the one with the best match. Sequences that are 25% longer or shorter than the median sequences were considered as fragments and excluded from the backbone. More specifically, PASTA first builds a starting tree, performs a tree-based clustering of the sequences, and builds a spanning tree from these clusters. Then it calls MAFFT v7.149b [162] to align the sequences in each cluster, and calls OPAL [353] to merge the alignments of adjacent clusters according to the spanning tree, and finally uses transitivity to perform the subsequent merging.

To ensure the quality of the alignment, we filtered out extremely gappy sites and sequences: sites with more than 90% gaps were deleted from the alignments, followed by the dropping of sequences with more than 66% gaps.

6.4.9 Filtering of marker genes

To ensure the quality of the species tree built upon these marker genes [298], we filtered out the genes that were not aligned reliably by UPP. As such, the marker genes with more than 75% gaps in the aforementioned alignments were excluded from the pool, leaving 381 marker genes in total. The threshold 75% was chosen based on the distribution pattern of per-gene alignment quality (Supplementary Fig. F.2d).

6.4.10 Filtering of outlier taxa from gene trees

We removed suspected outliers by detecting the taxa on disproportional long branches and filtering them out from the phylogeny inferred by FastTree [261]. To do this, we applied TreeShrink [199] v1.1.0, a method that simultaneously detects long branches on a set of gene trees by identifying a set of taxa that could be removed from each gene so that the gene trees are maximally reduced in diameter. We used FastTree 2.1.9 to infer preliminary gene trees of the 381 selected genes, then ran TreeShrink to detect outlier long branches in these trees, with the per-species test with $\alpha = 0.05$ (5% false-positive tolerance). Finally, we dropped genomes that contained less than 100 marker genes post gene tree filtering.

6.4.11 Gene tree reconstruction

Gene tree topologies were reconstructed using the maximum likelihood (ML) method as implemented in the state-of-the-art phylogenetic inference program RAxML 8.2.10 [321]. The best amino acid substitution model for each of the 381 universal marker genes was inferred using RAxML's built-in script ProteinModelSelection.pl. Three phylogenetic trees were reconstructed for each gene family: one using a starting tree computed by the fast ML approach implemented in FastTree) and the other two using parsimony trees built with random seeds 12345 and 23456. RAxML was executed with the ML search convergence criterion (-D) and the CAT rate heterogeneity model without final optimization (-F) to reduce the execution time.

For each of the 1,143 topologies (3 × 381), another RAxML run was executed to optimize branch lengths and to compute likelihood scores under the robust but expensive Gamma rate heterogeneity model. Because of numerical instability, at least one of the RAxML runs failed for 39 of the 381 gene families. For those cases, IQ-TREE 1.6.1 [231], an alternative and faster maximum likelihood program, was used instead to optimize branch lengths using the same model (G4). The tree with the highest likelihood score among the three runs was retained for downstream applications. In 161 gene families, this tree was from the run with the FastTree starting tree, while in the remaining gene families the best tree was from either one of the random seeds.

6.4.12 Species tree reconstruction by summarizing gene trees (ASTRAL)

A species tree was reconstructed by summarizing the 381 gene trees, using ASTRAL-MP [372] (implementing ASTRAL-III algorithm [378]) 5.12.6a. This analysis was run on the Comet supercomputing cluster using 24 cores and 4 GPU acceleration. In the resulting tree, the branch lengths represent the units of coalescence. Each branch has three support values: 1) effective number of genes (EN): number of gene trees that contain some quartets around that branch; 2) quartet score (QT): proportion of the quartets in the gene trees that support this branch; 3) local posterior probability (LPP): the probability this branch is the true branch given the set of gene trees (computed based on the quartet score and assuming incomplete lineage sorting (ILS)) [297].

6.4.13 Branch length estimation for the ASTRAL tree

The branch lengths of a summary tree generated by ASTRAL are in coalescent units and only for internal branches. In order to obtain “conventional” branch lengths, i.e., the expected number of amino acid substitutions per site, we ran IQ-TREE using the concatenated alignment (most conserved or randomly selected sites as described below) as input, the ASTRAL tree as the topological constraint, and the LG + Gamma model. Branch lengths obtained using both site categories were highly correlated (Supplementary Note F.1.2).

6.4.14 Species tree reconstruction based on the concatenated alignment (CONCAT)

The alignments of the 381 marker genes were concatenated into a supermatrix. Due to the computational challenge in running classical maximum likelihood tree reconstruction on the full-scale data set, we had to downsample it to around 38k amino acid sites. In order to explore the impact of site sampling on tree topology and branch lengths, we separately adopted two strategies for site sampling: 1) selected up to 100 most conserved sites per gene. The degree of conservation was estimated using the "trident" metric [344], which is a weighted composition of three functions: symbol diversity, stereochemical diversity, and gap distribution. The PFASUM60 substitution matrix was used for computing the stereochemical diversity [165]. 2) randomly selected 100 sites per gene from sites with less than 50% gaps.

For the downsampled supermatrix, a maximum likelihood tree was first built using FastTree, with LG model for amino acid substitution and Gamma model for rate heterogeneity. Using this FastTree tree as the starting tree, plus two maximum parsimony trees generated from random seeds (12345 and 23456), we performed three independent runs using RAxML, with the LG + CAT models (PROTCATLG), with rapid hillclimbing (-f D) and without final Gamma optimization (-F). With the resulting topologies, we performed branch length optimization and likelihood score calculation using IQ-TREE, with the LG + Gamma models (LG+G4). We further performed 100 rapid bootstraps using RAxML to provide branch support values.

6.4.15 Species tree reconstruction based on ribosomal proteins

To test the impact of choice of marker gene set on the topology and relative distances among major taxonomic groups, we conducted a separate analysis in which the species tree was built using ribosomal protein sequences. We identified and extracted 30 ribosomal protein families using the program PhyloSift 1.0.1 [63] with its marker database released on August 8, 2017. If more than one copy of a marker protein was detected in a genome, all copies were discarded. After this filtering, genomes with fewer than 25 marker proteins were dropped from the data set, resulting in a total of 9,814 genomes out of the original 10,575. Sequences of each ribosomal protein family were aligned using UPP as described above. The resulting alignments were concatenated and subjected to RAxML tree reconstruction using the LG model for amino acid substitution [181] (which is the best model for 304 out of the 381 genes based on RAxML's model selection) and the CAT model for rate heterogeneity (PROTCATLG). The resulting tree was then fed into IQ-TREE for branch length optimization, with the Gamma model for rate heterogeneity. One hundred rapid bootstraps were executed in RAxML to provide branch support.

The same concatenated alignment was also used to estimate the branch lengths for the ASTRAL tree based on the 381 marker gene trees. Because the quality of an ASTRAL tree improves as the number of gene trees increases (Supplementary Fig. F.10a, c), running ASTRAL on only 30 trees of structurally and functionally highly related genes is of limited value. Thus we decided not to run ASTRAL *de novo* but only to assess the impact of ribosomal proteins on the branch lengths of the existing ASTRAL tree.

6.4.16 Species tree reconstruction and branch length estimation with CPR taxa excluded

We followed the same protocol as stated above to reconstructed species trees and estimate branch lengths based on the protein sequence alignments with the 1,454 CPR taxa removed, leaving 9,121 taxa. Only one modification was made to the main protocol in order to reduce the computational expense for reconstructing the 381 gene trees: Instead of running RAxML three times per gene and selecting one tree with the highest Gamma likelihood, we ran RAxML once per gene using the random seed 12345. The two alternative site sampling schemes: most conserved ("cons") and randomly selected ("rand") as demonstrated in the main result were both tested, using the same amino acid sites as in the main protocol in each scheme.

6.4.17 Species tree reconstruction using site heterogeneous models (PMSF)

We built alternative CONCAT trees using the posterior mean site frequency (PMSF) method [348] implemented in IQ-TREE, which considers mixture classes of rates and substitution models (here the LG model) across sites. Because this method is computationally expensive, we downsampled the 10,575 taxa to 1,000 (see below for the taxon downsampling strategy). ModelFinder (which is part of IQ-TREE) [161] was used to select an optimal model among the empirical profile mixture models C10 to C60 [309], plus the site homogenous model (with Gamma rate across sites) as a control. This analysis consistently chose C60 as the optimal model for all tests. Therefore we used the LG+C60 model for PMSF phylogenetic reconstruction. PMSF requires a guide tree, which we obtained from ModelFinder results. Computational challenge limited this analysis to at most 1,000 taxa (which consumed 1.43 TB memory, close to the 1.5 TB physical memory equipped in our high-memory nodes). Branch support values were computed using the ultrafast bootstrap (UFBoot) [136] method implemented in IQ-TREE. In parallel to this analysis, we performed phylogenetic inference using the Gamma model (+G) or the FreeRate [316] model (+R) on the same 1,000-taxon input data for comparison purpose.

6.4.18 Species tree reconstruction using implicit methods

We applied two implicit strategies for inferring the evolutionary relationships among the sampled genomes. They are not based on the alignment of homologous features across multiple genomes, but instead, are based on the pre-defined distances among genomes. Specifically, they are the Jaccard distances defined by the MinHash signature (see above), and by the presence / absence of the 400 marker genes (see above). The conventional neighbor joining (NJ) method as implemented in ClearCut 1.0.9 [88] was used to reconstruct phylogenetic trees from the two distance matrices, respectively.

6.4.19 Rooting and post-manipulation of species trees

We rooted the species tree at the branch connecting the Archaea clade and Bacteria clade, according to the widely adopted hypothesis of life evolution [128, 102, 61]. The absence of Eukaryota does not impact the placement of root, since Eukaryota is considered derived, as a sister group or ingroup of Archaea in this hypothesis. We want to remind readers that this hypothesis is not without controversy [174, 52]. The discovery and study of CPR and other divergent or transitional groups may provide materials for a second examination of this hypothesis, although this is beyond the scope of this study.

Internal nodes were flipped to follow the descending order (i.e., child nodes are sorted from less descendants to more descendants). Incremental numbers were assigned to internal node IDs in a pre-order traversal of the tree starting from the root (i.e., root = N1, LCA of Archaea = N2, LCA of Bacteria = N3, etc.). These node IDs can be used as unique identifiers in downstream analyses and applications.

6.4.20 Phylogeny-based downsampling of taxa

We designed a protocol to downsample taxa from the 10,575 genomes for further phylogenetic analyses. We adopted the relative evolutionary divergence (RED) metric [246], as the core of our subsampling strategy. This metric allowed us to select large clades that best represent the deep phylogeny. Specifically, we calculated RED for all nodes (terminal and internal) of the ASTRAL tree (i.e., the tree shown in Fig. 6.1) using TreeNode functions implemented in scikit-bio 0.5.2 [302]. Nodes were selected iteratively from the low end of the RED list, with ancestral nodes (if any) of the current node dropped from the selection at each iteration, until the desired number of clades n was achieved.

Within each selected clade, four criteria were sequentially applied to the descendants until one taxon was selected: 1) contains the most marker genes; 2) contamination level is the lowest; 3) DNA quality score is the highest; 4) random selection (if there were still more than one taxon after applying the other three criteria). This protocol guaranteed the selection of *n* taxa which maximize the representation of deep phylogeny.

6.4.21 Visualization and annotation of trees

Unique colors were assigned to selected NCBI-defined taxonomic groups above phylum, and phyla with 100 or more representatives in the sampled genomes. Colors of taxa were directly assigned based on their NCBI taxonomy assignment. Colors of clades and branches were determined based on the tax2tree decoration. The trees were rendered using iTOL v4 [186] (unrooted or circular layouts) or FigTree 1.4.3 [1] (rectangular layout).

6.4.22 Comparison of multiple trees

We used both the classical Robinson–Foulds (RF) metric [277] (calculated using scikit-bio's "compare_rfd" function) and the quartet score (calculated using ASTRAL) to quantify the topological concordance between a pair of trees. Furthermore, we used the "tip distance" (TT), calculated using scikit-bio's "compare_tip_distances" function, to measure the correlation of the phylogenetic distances among taxa in a pair of trees. It equals $(1-r)/2$, where r is the Pearson correlation efficient between the tip-to-tip distance (i.e., total length of branches connecting two tips) matrices of the two trees. Because the two trees might have different sets of taxa, we first truncated them using the "shear" function implemented in scikit-bio so that they both only contained the shared taxa. This enabled the subsequent computation of the three metrics.

For a set of multiple trees (species trees or gene trees), a matrix of the pairwise RF distance, quartet distance (1 - quartet score) or tip distance was constructed, based on which subsequent statistical analyses were performed to assess the clustering pattern of trees, as stated below.

6.4.23 Clustering analysis of multiple trees

We used several statistical approaches to assess the clustering pattern of multiple trees based on the RF, tip or quartet distance matrices built as stated above:

1 Hierarchical clustering, using the "linkage" function implemented in SciPy 1.1.0 [155].

2 mMDS, as detailed above.

3 Principal coordinate analysis (PCoA), performed using QIIME 2's "pcoa" command, and visualized using Emperor. This method aims to visualize the biggest variance in a few dimensions, as compared to mMDS as explained above.

4 Permutational multivariate analysis of variance (PERMANOVA) [12], performed using QIIME 2's "beta-group-significance" command, with 999 permutations (the default setting). This method evaluates the statistical significance of grouping of trees by a certain variable such as method, site sampling and taxon sampling.

6.4.24 Cross-comparison of the ASTRAL and CONCAT trees

The first challenge for this comparison was that the branch support values were estimated using completely different methods (local posterior probability vs. rapid bootstrap) and so are not directly comparable. We manipulated the trees so that they have the same overall resolution: First, we collapsed the low-supported branches in the CONCAT tree (by conserved sites), using the commonly accepted bootstrap threshold: 50. This left 9,595 internal nodes. Then we performed branch collapsing to the ASTRAL tree, from the low end of the range of local posterior probability (lpp), until it reached 0.68057, also leaving 9,595 internal nodes.

The second challenge was that large-scale trees are difficult to align and to display. We collapsed the two trees so that they have 50 paired clades with at least 50 descendants each. For each pair of clades the descendants are identical. The remaining tips were pruned. This operation left 7,764 taxa in each tree. The sizes of the 50 chosen clades are 155.3 ± 106.9 (mean and standard deviation).

A tanglegram of the resulting collapsed trees was reconstructed using Dendroscope 3.5.9 [146]. In our case, the clades were fully-aligned. The tanglegram was then rendered back-to-back without the need for displaying the connector lines.

6.4.25 Calculation of the relative Archaea-Bacteria distance

We calculated the phylogenetic distance (sum of branch lengths) between every pair of taxa in a tree using scikit-bio 's "tip_tip_distances" function. The pairs were divided into three groups: A-A, A-B, and B-B (A and B are abbreviations for Archaea and Bacteria). Within each group, the mean distance was calculated. Then the overall relative A-B distance was calculated as: mean(A-B)2 / (mean(A-A) × mean(B-B)). Note that due to HGT and other reasons, archaeal and bacterial taxa are rarely perfectly separated in individual gene trees. Therefore the calculated distance should be interpreted as the average evolutionary distance between archaeal and bacterial genomes, instead of the distance between the two clades.

6.4.26 Test for amino acid substitution saturation

We followed the principle introduced by Jeffroy *et al.* [151] to test for the saturation. Specifically, we wanted to test whether the degree of saturation on inter-domain taxon pairs (Bacteria vs. Archaea) is larger than that on intra-domain pairs. For each domain, 100 taxa were randomly sampled for this analysis. We plotted the phylogenetic distance, i.e., the sum of branch lengths between two tips, as the *x*-axis, versus the Hamming distance of gap-free sites per each alignment between a pair of sequences, as the *y*-axis (Supplementary Fig. F.19a-d). Because the three categories of taxon pairs have differential distribution on the *x*-axis, we further binned on the x-axis and performed comparison within each bin (Supplementary Fig. F.19e, f).

6.4.27 Phylogenetic analysis with latest genome availability

We made several modifications to the main protocol to reduce the computational expense for this rapid test of the extended set of 10,762 (10,575 + 187) genomes: UPP was called in "insertion" mode to update the existing amino acid sequence alignments. In-house scripts were used to locate the same set of sites instead of performing *de novo* site sampling. Both ASTRAL and CONCAT methods were used to build species trees. For CONCAT, we used IQ-TREE in "fast" mode to build *de novo* species trees from concatenated alignments without using a predefined starting tree. For ASTRAL, we kept the same analysis parameters to build a species tree from the 381 gene trees, whereas the gene trees were built as follows to save computation while maintaining high quality:

Firstly we used the previous gene trees as topological constraints (-g) to incorporate the new taxa using RAxML. Then we used those trees as starting trees (-t) to perform *de novo* ML searches using RAxML. This way, we only did *de novo* ML search once instead of three as previously, but we argued that the generated gene trees would have comparable ML score as in the previous procedure. To test this hypothesis, we randomly selected 10 genes to generate four trees each: (1) RAxML with FastTree tree as starting tree; (2) & (3) RAxML with random starting trees with two different random seeds; (4) RAxML tree generated using the described procedure. Note that the tree having highest likelihood score among (1), (2), and (3) defines the ML tree in the previous procedure. Our results showed that the gene trees generated by (4) have higher likelihood scores than the best of (1), (2), and (3) in six of 10 of the tested genes. Besides, we use a χ^2 test to show that the trees (4) have higher chance to be the best tree than (1), (2), and (3). In this test, the null hypothesis H0 is that (4) has the same chance to be the best tree among the four trees. Applying the test on the 10 selected genes, we rejected H0 with p-value = 0.011.

6.4.28 Divergence time estimation using maximum likelihood

We used the maximum likelihood tool r8s 1.81 [294] to estimate the divergence times based on the species trees. Specifically, we used the Langley-Fitch (LF) method [177] , which assumes a universal molecular clock (substitution rate) for the entire tree, with the truncated-Newton (TN) method for optimizing the likelihoods of branch lengths [228]. A recent study showed that this method has comparable estimation accuracy when benchmarked against the more sophisticated Bayesian framework, but its computation is significantly faster [340], thus suitable for the size of our dataset. Near-zero branches were collapsed to avoid numerical errors. Ten replicates with random initial conditions were performed for each setting. In each replicate, three restarts were executed after the initial optimization with a random perturbation factor of 5%. Replicates that failed to pass the gradient check were discarded. The divergence times estimated by the run with the highest likelihood score and the mean and standard deviation of those by all successful runs were reported.

6.4.29 Divergence time estimation using Bayesian inference

We used the Bayesian tool BEAST 1.10.4 [326] to estimate divergence times. Considering the computational expense, we randomly selected 5,000 amino acid sites from the full-length alignment, and downsampled the original 10,575 taxa to 100. Taxon sampling was performed using the same RED-guided protocol (see above), but was manually modified afterwards to ensure sufficient sampling around the calibration point. Two alternative molecular clock models were used: the strict clock model, or the uncorrelated relaxed clock model with a lognormal distribution (UCLD) [77]. The species tree was modeled using a Yule process [108], with topology fixed to the ASTRAL tree. Logs of MCMC runs were examined using Tracer 1.7.1 [269]. Burn-ins were set to be at least 10% of iterations, or higher depending on the manual observation of traces. Sufficient MCMC iterations were executed to ensure that the effective sample size (ESS) of the reported parameters was at least 150.

6.4.30 Tree-based taxonomic curation and annotation

We used the program tax2tree (commit 99f19be) [209] to curate the original NCBI taxonomy [90] assignment of genomes based on the phylogenetic trees and to annotate the internal nodes of the tree using most appropriate taxonomic labels. The same program was used to curate multiple databases such as the classical Greengenes [209] and the recent GTDB [246]. The program took as input the species tree and the original NCBI taxonomy and inferred the most plausible taxonomic annotation at every node of the tree, as determined using an F-measure scoring system across candidate taxonomic terms. In scenarios where one term was estimated to be the best candidate for multiple, independent clades (i.e., para/polyphyly), a numeric suffix was appended to the term to indicate the grouping and order (from more descendants to less) of those clades. For example, Firmicutes_1 is the largest clade assigned to the paraphyletic phylum Firmicutes, followed by Firmicutes_2, Firmicutes_3, etc. Based on the decorated tree, correct taxonomic names were re-generated for unclassified and mis-annotated genomes. Taxonomic groups represented by only one genome in this work were back-filled post tax2tree annotation.

6.4.31 Assessment of cladistic properties of taxon sets

The cladistic property of a taxonomic group (or an arbitrarily defined taxon set) with reference to a species tree was evaluated using three methods:

1) The strict definition of “monophyly”: when a clade contains all genomes assigned to a single taxonomic group and no other genomes, this taxonomic group is considered monophyletic. Further, we identified “relaxed” monophyletic groups compared to the aforementioned “strict” scenario. In the “relaxed” scenario, if a clade consists of genomes assigned to a taxonomic group, and genomes without assignments at the same taxonomic rank (i.e., unclassified), this taxonomic group is still considered monophyletic.

2) tax2tree’s classification consistency score, representing the fraction of tips within that clade relative to the total number of tips in the tree which are of that taxon. Consistency = 1 is equivalent to strict monophyly.

3) The ASTRAL-computed quartet score of this taxonomic group, i.e., the fraction of quartets in the tree that supports this taxonomic group as monophyletic, i.e., separates this taxonomic group from the others.

4) An approach introduced in DiscoVista [299] which evaluates and visualizes the compatibility between a given taxon set and a tree with branch support values. It computes a “support” or “rejection” degree as follows: If the taxon set constitutes a monophyletic clade in the tree, it is supported; and the support degree (green) is the support value of the branch connecting the lowest common ancestor of the clade to its parent. On the other hand, if it is not a monophyletic group in the original tree, but after contracting branches with support values below a threshold, the monophyly can no longer be rejected due to polytomy, the lowest threshold is considered the rejection degree (with a negative sign) (magentar).

6.4.32 Evaluation of GTDB taxonomic groups

We downloaded GTDB [246] release: 86.1 from http://gtdb.ecogenomic.org/. The format of genome identifier in GTDB was matched to that of our work (e.g., GB_GCA_000123456.1 was translated into G000123456). Following the protocols described above, we evaluated the GTDB phylogeny and taxonomic units, and annotated our species trees using the GTDB taxonomy.

6.4.33 Statistics

Statistical analyses and plotting were performed using Python 3.6 and QIIME 2 release 2017.12. Specifically, PERMANOVA test was performed using QIIME 2's "beta-group-significance" command. Independent or paired two-sample t-test was performed using scipy 1.1.0's "ttest_ind" and "ttest_rel" commands, respectively. Fisher's exact test was performed using scipy's "fisher_exact" function. Linear regressions were performed using scipy's "linregress" function. The *p*-value was computed using a two-sided Wald test in which the null hypothesis was slope = 0. Gaussian kernel density estimations were performed using seaborn 0.9.0's "distplot" function. Hierarchical clustering was performed using scipy's "linkage" function. Quantile-quantile (Q-Q) plot was computed using scipy's "probplot" command. Redundancy analysis (RDA) was performed using vegan 2.4.4's "rda" and "ordiR2step"commands. Dimension reductions were performed using mMDS implemented in scikit-learn 0.19.2, or PCoA implemented in QIIME 2 (both detailed above). Pairwise distances based on k-mer signatures and on marker gene presence/absence were computed using the Jaccard index (see above). Branch supports in the phylogenetic trees were computed using rapid bootstrap implemented in RAxML 8.2.10, and ultrafast bootstrap implemented in IQ-TREE 1.6.1, and local posterior probability implemented in ASTRAL 5.12.6a (detailed above). Robinson-Foulds (RF) distance and "tip distance" were calculated using scikit-bio 0.5.2. Quartet scores were calculated using ASTRAL.

6.4.34 Data availability

The datasets generated during and analyzed during the current study are publicly available at GitHub: HYPERLINK ”https://github.com/biocore/wol” (DOI: 10.5281/zenodo.3524546), under the BSD 3-Clause license. All relevant data are available from the authors. The source data underlying Figs. 6.1-6.6, and Supplementary Figs. F.1-F.27 are provided as a source data files.

6.4.35 Code availability

The Python implementations of the prototype selection algorithms for genome subsampling are publicly available at: https://github.com/biocore/wol/ (DOI: 10.5281/zenodo.3524546), under the BSD 3-Clause license. A copy of the code is provided in Supplementary Software.

6.5 Acknowledgements

We are grateful to Victor Nizet, Eric Allen, Milton Saier, Jürgen Schulze, Toni Gabaldón Estevan, Benjamin Hillmann, Arturo Medrano-Soto, Madhusudan Gujral, Monika Kumaraswamy, Jeff Dereus, Anupriya Tripathi, Amnon Amir, Serene Jiang, Yoshiki Vázquez-Baeza, Thant H. Zaw for insightful discussions on this study and additional assistance. **Funding**: This work used the Comet supercomputer at the San Diego Supercomputer Center through allocation BIO150043, awarded to R. K., L. S., and S. M. through the Extreme Science and Engineering Discovery Environment (XSEDE). Both Comet and XSEDE are funded by the National Science Foundation.

6.6 Authors contributions

Q.Z. and R.K. conceived the study; Q.Z. and U.M. led data analysis; U.M. and S.J. led algorithm development and implementation; W.P., U.M. and F.A. led high-performance computing; Q.Z. led result interpretation and manuscript writing; Q.Z., G.A.A., J.B., Z.W., E.S., M.R., K.C., Y.Y. and S.M. contributed to algorithm development and implementation; U.M., S.J., J.G.S., P.B., D.M., S.H., N.S., J.J., S.P., C.H., S.M. and R.K. contributed to result interpretation; F.A., T.K., J.T.M. and S.M. contributed to data analysis; Q.Z., F.A., E.K. and Z.Z.X. contributed to data collection; J.G.S., D.M., D.K., W.L., N.S., L.S., S.M. and R.K. contributed to study design; all coauthors contributed to the writing and discussion of the manuscript.

6.7 Competing interests

All authors declare that they have no competing interests. Materials & Correspondence Correspondence and material requests should be addressed to Rob Knight (robknight@ucsd.edu).

6.8 Figures

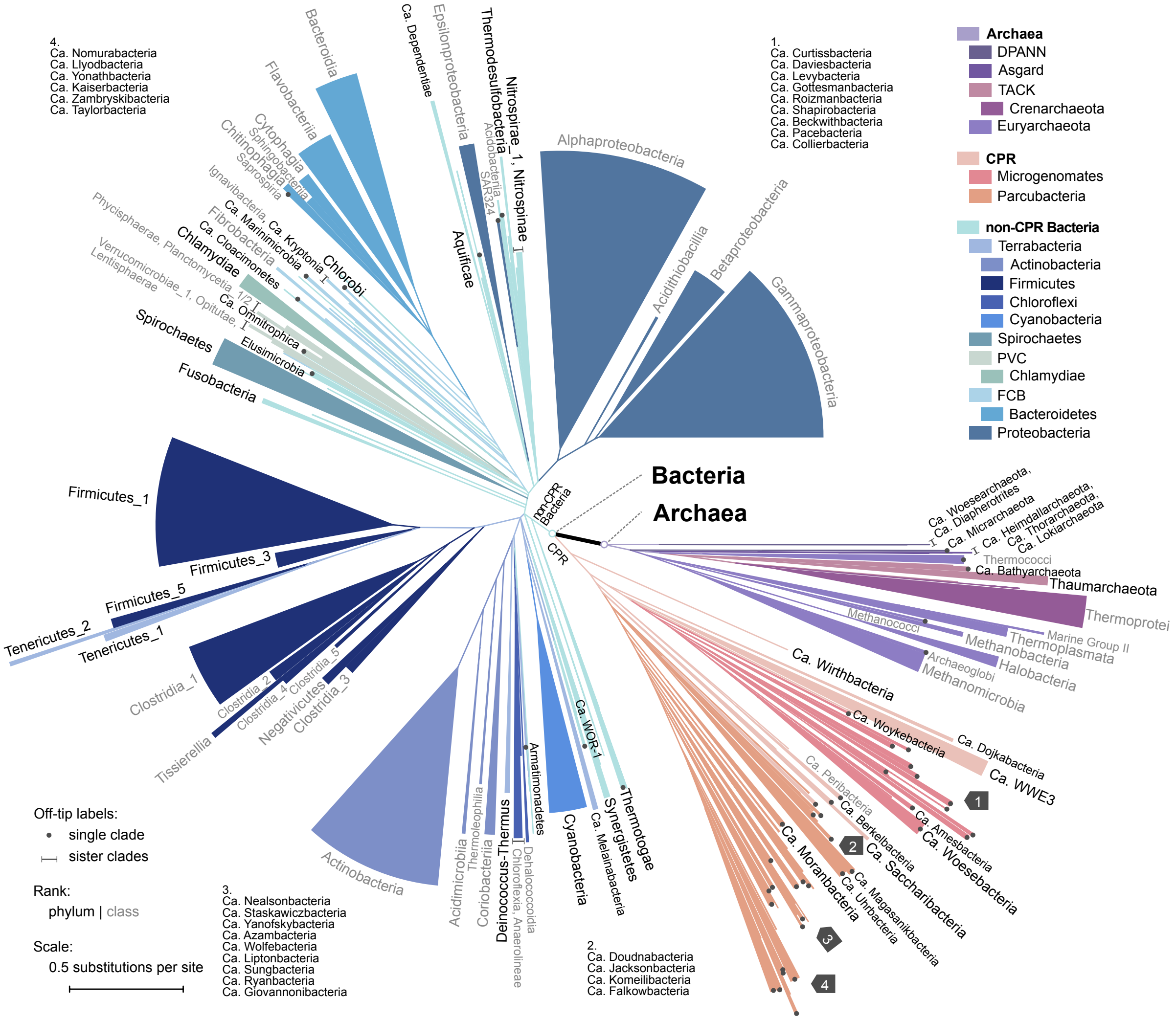


Figure 6.1: A new view of the bacterial and archaeal tree of life. The tree contains 10,575 evenly distributed bacterial and archaeal genomes, with topology reconstructed using ASTRAL based on individual trees of 381 globally sampled marker genes, and branch lengths estimated based on 100 most conserved sites per gene. Branches with effective number of genes (en) ≤ 5 and local posterior probability (lpp) ≤ 0.5 were collapsed into polytomies. Taxonomic labels at internal nodes and tips reflect the tax2tree curation result. Color codes were assigned to above-phylum groups and phyla with 100 or more representatives. To display the tree in a page, it was collapsed to clades (sectors) representing phyla with at least one taxon (black), and classes with at least 10 taxa (grey). The radius of a sector indicates the median distance to all descending taxa of the clade, and the angle is proportional to the number of descendants. For polyphyletic taxonomic groups, minor clades with less than 5% descendants of that of the most specious clade were omitted, while the remaining clades were appended a numeric suffix sorted by the number of descendants from high to low. Dots (single clade) and lines (sister clades) are used to assist visual connection between tips and labels where necessary. In four instances where visual space is inadequate (marked by grey arrows), groups of labels in clockwise order are provided in remote blank areas. Source data are provided as a Source Data file.

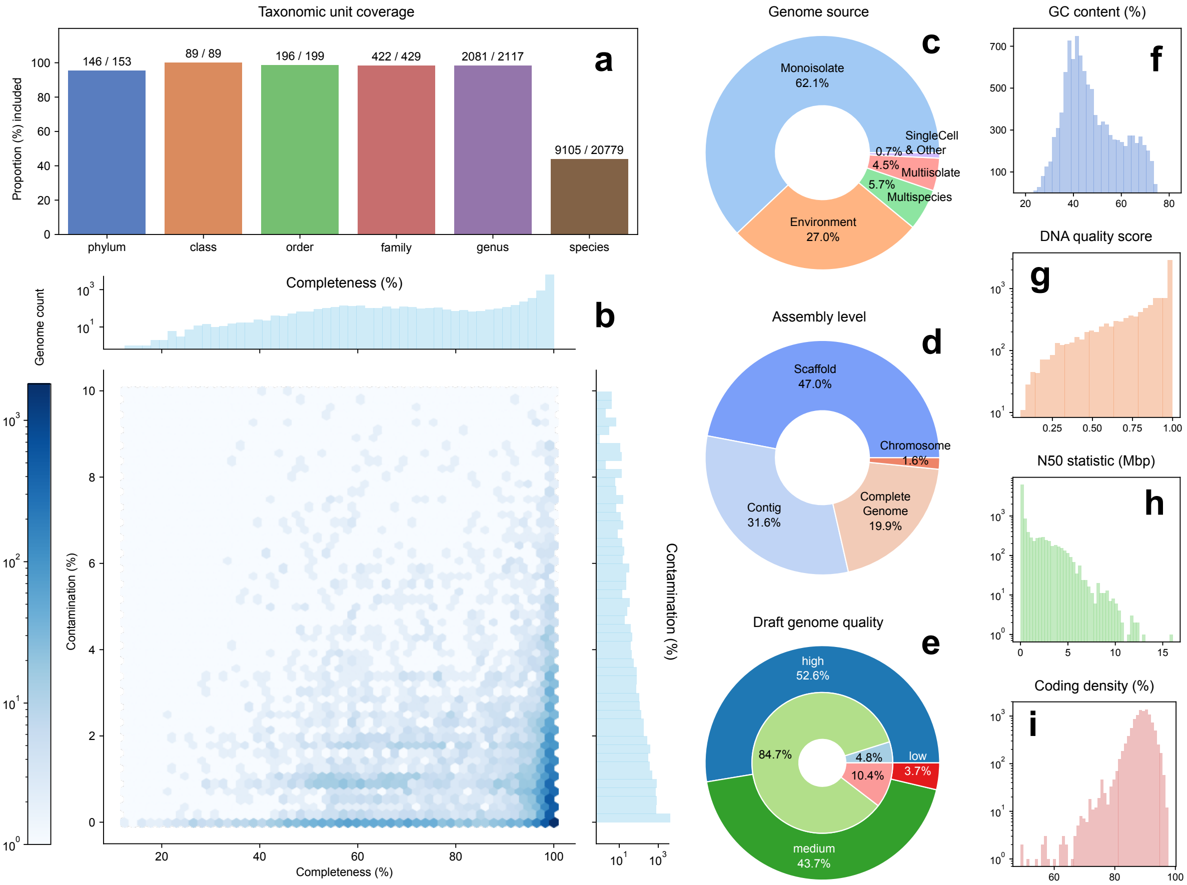
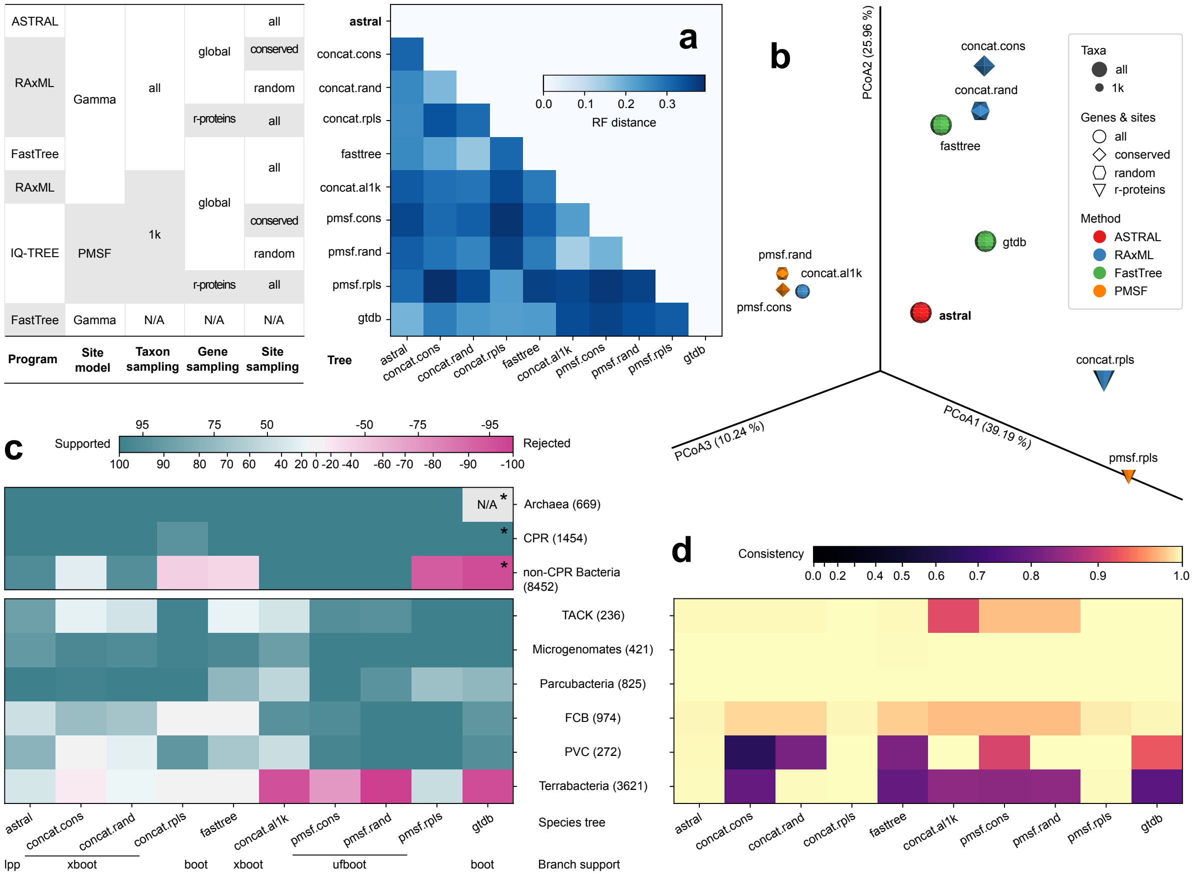


Figure 6.2: Statistics of the 10,575 bacterial and archaeal genomes selected for phylogenetic reconstruction. **a**. Numbers and proportions of NCBI-defined taxonomic units included from all 86,200 available genomes. **b**. Distribution of completeness vs. contamination scores computed by CheckM [247]. **c**. Distribution of genome sources, i.e., the "scope" property defined by NCBI. **d**. Distribution of genome assembly levels. **e**. Distribution of draft genome quality, determined following the standard established by Bowers *et al.* [34]. Specifically: "high": completeness $>$ 90%, contamination $<$ 5%, presence of 23S, 16S, 5S rRNAs and *geq* 18 tRNAs.; "medium": completeness $\geq$ 50%, contamination $<$ 10%; "low": completeness $<$ 50%, contamination $<$ 10%. The outer, darker circle represents all genomes. The inner, lighter circle represents genomes assembled from metagenomes (MAGs). **f**. Distribution of GC contents. **g**. Distribution of DNA quality scores, calculated following Land *et al.* [176]. **h**. Distribution of N50 statistics of nucleotide sequences per genome. **i**. Distribution of coding density. The *y*-axes in **f**-**h** represent genome counts. Source data are provided as a Source Data file.



Program	Site model	Taxon sampling	Gene sampling	Site sampling	Tree
ASTRAL	Gamma	all	global	all	astral
RAxML				conserved	concat.cons
				random	concat.rand
			r-proteins	all	concat.rpls
FastTree			global	all	fasttree
RAxML		1k			concat.al1k
IQ-TREE	PMSF			conserved	pmsf.cons
				random	pmsf.rand
			r-proteins	all	pmsf.rpls
FastTree	Gamma	N/A	N/A	N/A	gtdb

Figure 6.3: Comparison of topologies of multiple species trees. Nine species trees reconstructed in this study, plus a previously published tree, GTDB release 86.1 were cross-compared. The methods for building those trees were summarized in the inset table. **a**. Matrix of normalized Robinson–Foulds (RF) distance, which measures the overall topological discrepancy between two trees based on the shared taxa. **b**. PCoA of the RF distance matrix. **c**. Branch support-informed support/rejection degrees (see Methods) for the monophyly of Archaea, CPR and non-CPR Bacteria, and six super phyla defined by NCBI. Clades were defined according to the tax2tree auto-curation result based on each tree. Note that super phyla Asgard (eight taxa) and DPANN (five taxa) are not shown due to low taxon representativeness, but are discussed in Supplementary Note F.1.5. Numbers in parentheses indicate the number of taxa under each clade according to the ASTRAL tree. The types of branch support values are indicated below tree names: "lpp": local posterior probability, "boot": classical bootstrap, "xboot": rapid bootstrap, "ufboot": ultrafast bootstrap. Note that different branch support types cannot be directly compared. Note (*) that the GTDB phylogeny consists of two trees independently built for Archaea and Bacteria, respectively, thus the support/rejection for Archaea cannot be assessed. Meanwhile, the bacterial tree was rooted using a midpoint strategy, making Chloroflexi the outgroup to all remaining bacteria (including CPR). But it should not be considered as a rejection to the monophyly of non-CPR Bacteria in the sense of evolutionary relationships. **d**. Consistency between NCBI-defined super phyla and tree topologies, computed using tax2tree (see Methods). Source data are provided as a Source Data file.

6.9 Acknowledgements

Chapter 6, in full, is a reprint of the material as it appears in “Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea”. Zhu, Qiyun, Uyen Mai, Wayne Pfeiffer, Stefan Janssen, Francesco Asnicar, Jon G. Sanders, Pedro Belda-Ferre, Gabriel A. Al-Ghalith, Evguenia Kopylova, Daniel McDonald, Tomasz Kosciolek, John Yin, Shi Huang, Nimaichand Salam, Jian-yu Jiao, Zijun Wu, Zhenjiang Z. Xu, and Kalen Cantrell, Yimeng Yang, Erfan Sayyari, Maryam Rabiee, James T Morton, Sheila Podell, Dan Knights, Wen-jun Li, Curtis Huttenhower, Nicola Segata, Larry Smarr, Siavash Mirarab, Rob Knight. Nature communications 10, no. 1 (2019): 1-14. The dissertation author was a primary investigator and co-first author of this paper.

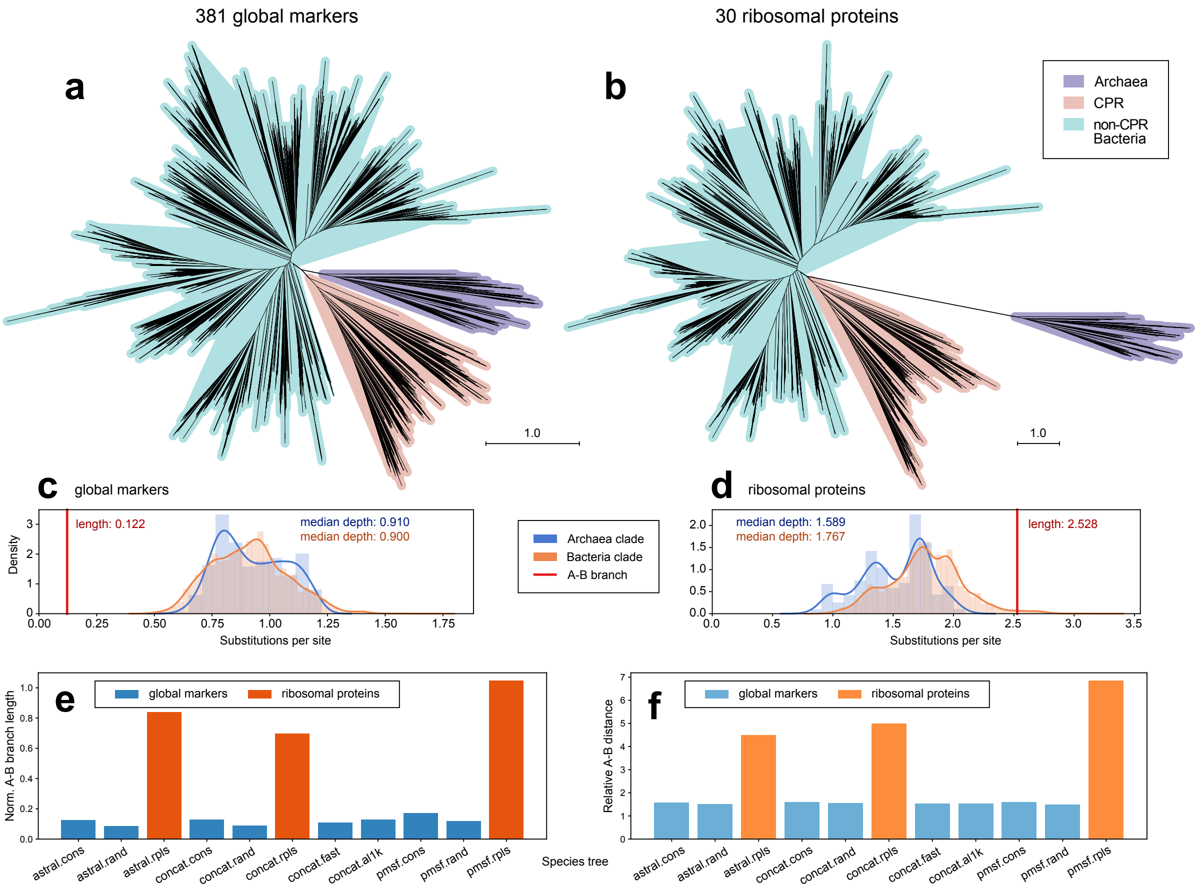


Figure 6.4: Evolutionary proximity between domains Archaea and Bacteria. **a** and **b**. The unrooted, drawn-to-scale ASTRAL tree with branch lengths estimated using the 381 global marker genes (conserved site sampling) (**a**) or using the 30 ribosomal proteins (**b**) are displayed, with color codes highlighting domain-level relationships (Archaea and Bacteria, the latter of which consisting of CPR and non-CPR Bacteria). Scales are in the unit of number of substitutions per site. c and d. For each domain (blue: Archaea, orange: Bacteria), the histogram and Gaussian kernel density function of the depths of all descendants (sums of branch lengths from a tip to the lowest common ancestor (LCA) of the clade) are plotted, with the median depth displayed; the length of branch connecting the LCA of A(rchaea) and the LCA of B(acteria) is marked as a red vertical line, with its value displayed. **e** and **f**. Evolutionary distance between domains Archaea and Bacteria by multiple trees (tree names follow Fig. 6.3), among which the branch lengths of the ASTRAL tree were estimated using differential gene and site samplings, respectively. Two metrics are displayed: the A-B branch length normalized by the median depth of all tips in the tree (**e**), and the relative distance between Archaea taxa (tips) and Bacteria taxa (**f**, see Methods). Color codes highlight evolutionary distances indicated by differential gene samplings (blue: the 381 global marker genes, orange: the 30 ribosomal proteins). Source data are provided as a Source Data file.

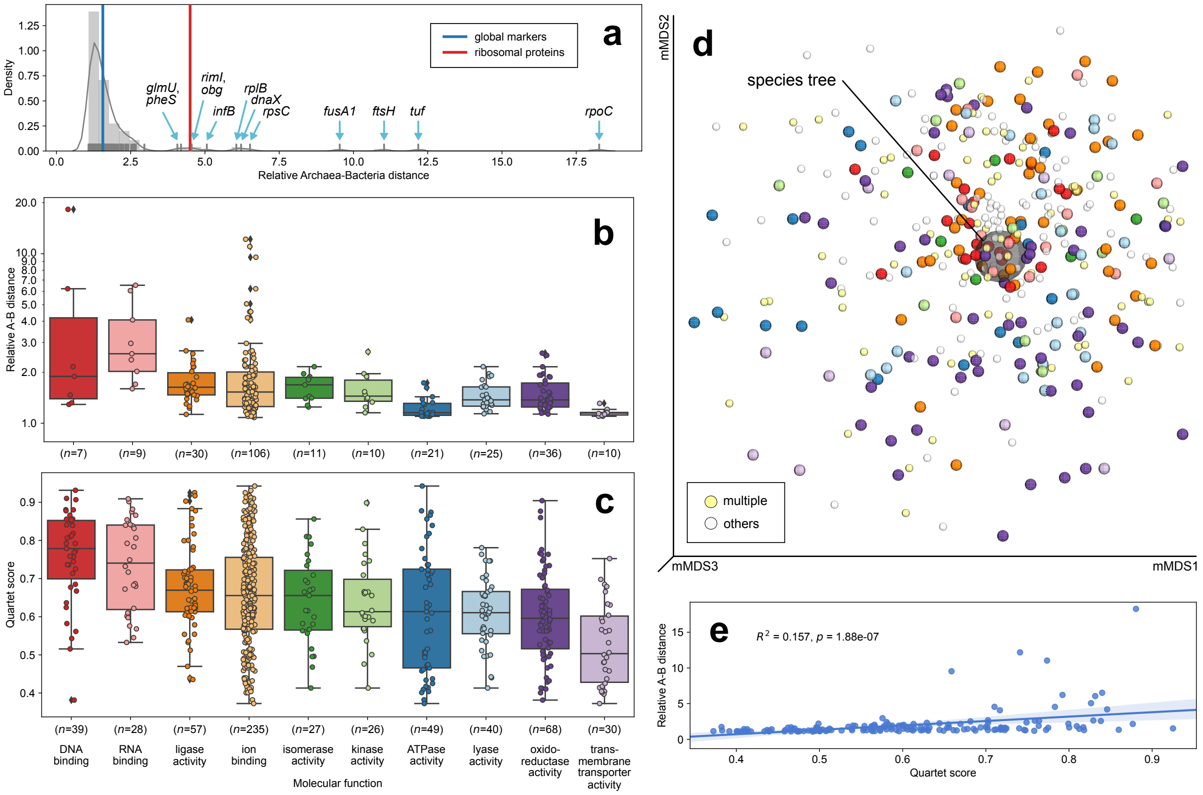


Figure 6.5: Relative Archaea-Bacteria distances indicated by individual gene trees and their concordance with the species tree. **a**. Distribution of relative A(rchaea)-B(acteria) distances of individual gene trees. A total of 161 gene trees are shown, selected such that both domains have at least 50% taxa represented in each tree. A histogram with Gaussian kernel density function and a rug plot representing individual data points are displayed. The blue and red vertical lines indicate the values of the ASTRAL species tree with branch lengths estimated using the global markers and the ribosomal proteins, respectively. Gene names are labeled at data points separated from the main cluster. **b**. Distribution of relative A-B distances by functional category (GO slim term under the "molecular function" master category) of the 161 gene trees. The top ten most frequently assigned categories in all gene trees are shown. Boxplot components: center line, median; box limits, upper and lower quartiles; whiskers, 1.5 × interquartile range; black diamonds, outliers. **c**. Distribution of quartet scores of all 381 gene trees vs. the species tree. Boxplot components are identical to **b**. **d**. mMDS plot of the quartet distance matrix of all 381 gene trees plus the species tree (semi-transparent big grey ball in the center). The color scheme for genes annotated by exactly one of the top ten functional categories (normal-sized balls) is consistent with **b** and **c**, except that the category "ion binding" is omitted due to its high frequency. Genes annotated by more than one of the top ten categories (light yellow), or by categories other than the top ten (semi-transparent light grey) are indicated by smaller balls. The ASTRAL species tree is represented by a large, grey ball. **e**. Linear regression of relative A-B distances vs. quartet scores of the 161 gene trees. The squared Pearson correlation coefficient (R^2) and two-tailed p-value are displayed. Source data are provided as a Source Data file.

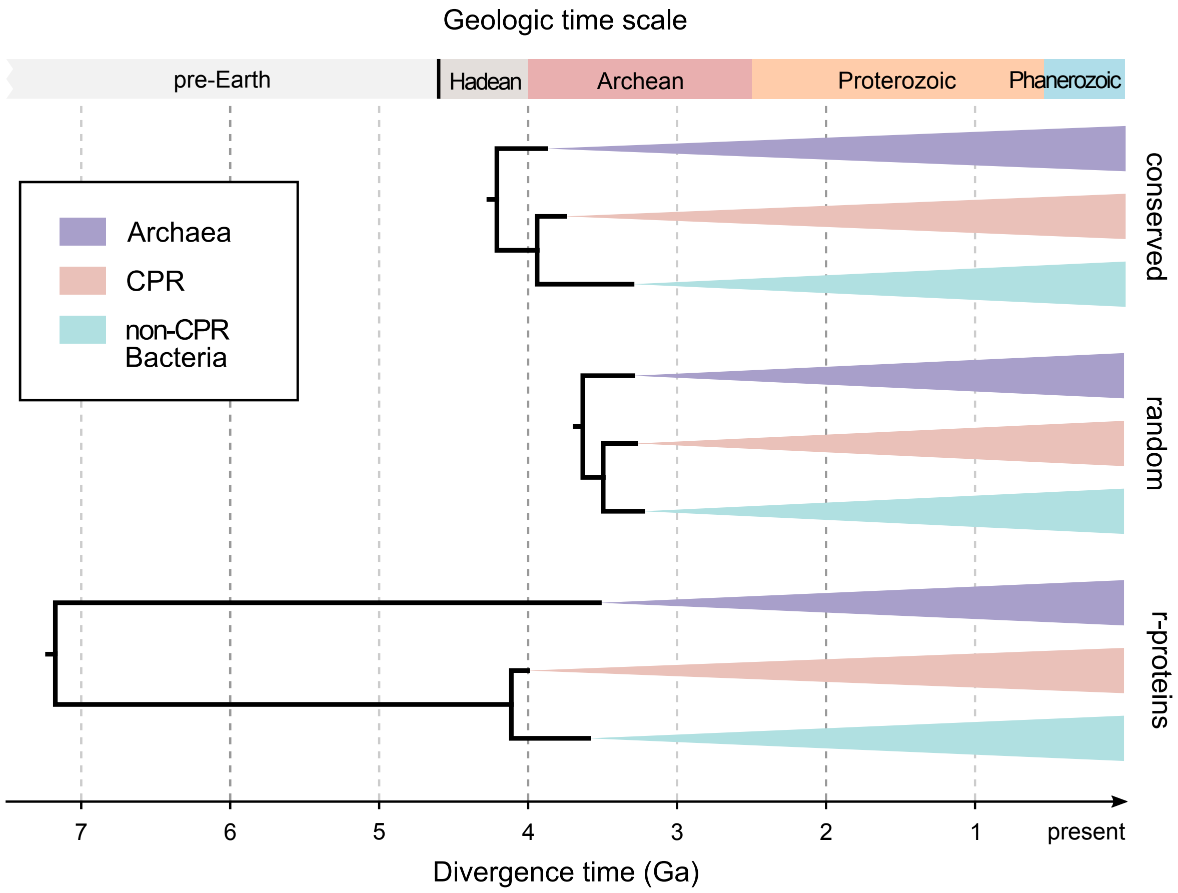


Figure 6.6: Estimated ages of basal diversifications. Divergence time estimation was performed using maximum likelihood based on the ASTRAL tree topology with branch lengths estimated using most conserved or randomly selected sites from the 381 global marker genes, or using the 30 ribosomal proteins. A universal clock was assumed. The Cyanobacteria/Melainabacteria split was constrained to 2.5-2.6 Ga. The best estimates from ten technical replicates per setting are displayed (see Supplementary Table F.8). Details and alternative results of divergence time estimation are elaborated in Supplementary Note F.1.6. Source data are provided as a Source Data file.

Appendix A

Supplementary materials for Chapter 1

A.1 Theorem proofs

A.1.1 Proof of Proposition 1

We start with a Lemma:

Lemma 1. *If (a,b) is a diameter pair of t, then for any $c,d \in \mathcal{L} - \{a\}$,*

$$max(\delta(c,b), \delta(d,b)) \geq \delta(c,d) \ .$$

Proof. Consider the quartet formed by the 4 leaves a,b,c,d in t.

Case 1: (Figure A.1c) a and b are on the same side of the quartet:

$$\delta(a,b) \geq \delta(a,d) \implies \delta(m,b) \geq \delta(m,d) \implies$$

$$\delta(c,m) + \delta(m,b) \geq \delta(c,m) + \delta(m,d) \implies \delta(c,b) \geq \delta(c,d)$$

Case 2: (Figure A.1d) Without loss of generality, we assume $\delta(n,c) \geq \delta(n,d)$. We will prove that $\delta(b,c) \geq \delta(c,d)$.

We have:

$$\delta(a,b) \geq \delta(a,c) \implies \delta(b,m) \geq \delta(c,m)$$

$$\implies \delta(b,m) + \delta(n,c) \geq \delta(c,m) + \delta(n,c)$$

$$\implies \delta(b,m) + \delta(n,c) \geq 2\delta(n,c)$$

$$\implies \delta(b,c) = \delta(b,m) + \delta(n,c) + \delta(m,n) \geq \delta(n,c) + \delta(n,d) = \delta(c,d)\ .$$

□

We now provide the proof of Proposition 1.

Proof. Consider an arbitrary leaf $b \in \mathcal{D}(t) - \{a\}$. We prove that $b \in \mathcal{D}(t\backslash_a)$.

Case 1: $(a,b) \notin \mathcal{P}(t)$. Because $b \in \mathcal{D}(t)$, there exists $c \in \mathcal{D}(t) - \{a\}$ such that (c,b) is a diameter pair of $t\backslash_a$. Therefore, $b \in \mathcal{D}(t\backslash_a)$.

Case 2: $(a,b) \in \mathcal{P}(t)$. Let (c,d) be a diameter-pair of $t\backslash_a$. According to Lemma 1, $max(\delta(c,b),\delta(d,b)) \geq \delta(c,d)$. Therefore, either (c,b) or (d,b) is a diameter-pair of $t\backslash_a$. Thus, $b \in \mathcal{D}(t\backslash_a)$. □

A.1.2 Proof of Theorem 1

Proof. We need to prove that a $\mathcal{R}_k(t)$ is either a reasonable removing set or it is not an optimal removing set. We proceed by contradiction. Assume $\mathcal{R}_k(t)$ is optimal but not a reasonable removing set. Let $\mathcal{R}_m(t)$ be the largest reasonable removing set that is a subset of $\mathcal{R}_k(t)$ (note $0 \leq m \leq k$). If $m = k$, then $\mathcal{R}_k(t)$ is a reasonable set, contradicting the assumption. For $m < k$, consider the tree $t\restriction_{\mathcal{R}_m(t)}$ and let a_m, b_m be its diameter pair. if $a_m \in \mathcal{R}_k(t)$ or $b_m \in \mathcal{R}_k(t)$, adding them to $\mathcal{R}_m(t)$ would generate a reasonable chain of size $m+1$, contradicting our assumption. If $a_m \notin \mathcal{R}_k(t)$ and $b_m \notin \mathcal{R}_k(t)$, all removals after m in $\mathcal{R}_k(t)$ fail to reduce the diameter, but removing either a_m or b_m would reduce the diameter. Thus, $\mathcal{R}_k(t)$ cannot be optimal, contradicting our assumption. □

A.1.3 Proof of Theorem 2

Proof. To remove k leaves from a singly paired tree t that has (a,b) as a diameter pair, at least one of a or b has to be removed (or else the diameter never decreases). Thus, three types of reasonable chains exist: those that contain only a, those that contain only b, and those that contain both a and b. Note that after removing a, by Proposition 1, removing b is a reasonable removal (and vice versa), and thus, removing both a and b is always reasonable (in either order).

Case 1: $a \in \mathcal{R}_{k-2}, b \notin \mathcal{R}_{k-2}$: If a reasonable chain has a but not b, by Proposition 1, b is in the diameter set at each step of the chain. Since b by definition is never removed and recalling that the tree is singly paired, at each step, there is only one reasonable removal (whatever leaf is on-diameter in addition to b). Therefore, only one reasonable chain does not include b.

Case 2: $a \notin \mathcal{R}_{k-2}, b \in \mathcal{R}_{k-2}$: Similar to Case 1, one such chain exists.

Case 3: $a \in \mathcal{R}_{k-2}, b \in \mathcal{R}_{k-2}$: In this case, the reasonable removing chain must start with a,b or b,a. In either ordering, we are left with the same induced tree, and need to remove $k-2$ more leaves. Therefore, the set of all reasonable removing sets in this case is: $\{(\{a,b\} \cup R)$ for R in $\mathcal{S}_{k-2}(t \upharpoonright_{\mathcal{L}-\{a,b\}})\}$.

Combining the three cases together, we have:

$$|\mathcal{S}_k(t)| = |\mathcal{S}_{k-2}(t \upharpoonright_{\mathcal{L}-\{a,b\}})| + 2$$

Let $s_k = |\mathcal{S}_k(t)|$. We have the following recursion:

$$s_k = \begin{cases} 1, & k = 0 \\ 2, & k = 1 \\ s_{k-2} + 2 & k \geq 2 \end{cases} \tag{A.1}$$

Thus, $s_k = k + 1$.

□

A.1.4 Proof of Proposition 2

Proof. Recall that the record of each internal node keeps track of the most distant leaves below two children of the node. When we remove a, only those nodes on the path from a to the root can have a change in their record. The first traversal of Algorithm 1 updates the records for those nodes, using simple recursive functions that can be computed in $O(1)$ per node.

According to Proposition 1, $b \in \mathcal{D}(t\backslash_a)$. Therefore, one of the longest paths in $t\backslash_a$ must include b; let c be the other leaf. The record of the LCA of c and b, after the update in the first round, will have the value of this longest value. Thus, by checking the updated record for all nodes in the path from b to the root we will find the maximum value. Moreover, when updating the records in the first traversal from a to the root, we have already checked all the nodes from the $LCA(a,b)$ to the root. In the second traversal, we check the nodes from b to $LCA(a,b)$, completing the search. Each of the two traversals of Algorithm 1 visits at most h nodes and only need constant time operations in each visit. Therefore, the overall time complexity of Algorithm 1 is $O(h)$. □

A.1.5 Proof of Theorem 3

First, we prove the following lemmas:

Lemma 2. *All the longest paths in any tree have the same midpoint.*

Proof. If t has only one diameter pair, then Lemma 2 is trivially correct.

If t has more than one diameter pair, let (a,b) and (c,d) be two distinct diameter pairs of t and let m be the midpoint of the path between a and b. We prove that m is also the midpoint of the path between c and d, that is m lies on that path between c and d and $\delta(m,c) = \delta(m,d)$. w.l.o.g, we suppose $\delta(m,c) \geq \delta(m,d)$.

- We prove that the path between c and d must pass m; that is, c and d belong to two different subsets in the partition defined by m on $\mathcal{L}$ (we call elements of the partition a "side"). We prove by contradiction, assuming c and d belong to the same side of m. Then $\delta(m,c) + \delta(m,d) > \delta(c,d)$. Also, either a or b must be on a different side from c and d to m (by definition, a and b cannot be on the same side to m). Suppose a is in a different side from c and d to m. Then: $\delta(a,b) \geq \delta(a,c) \implies \delta(a,m) + \delta(m,b) \geq \delta(a,m) + \delta(m,c) \implies \delta(m,b) \geq \delta(m,c) \implies \delta(m,a) \geq \delta(m,c)$. So we have,

  $\delta(a,c) = \delta(m,a) + \delta(m,c) \geq 2\delta(m,c) \geq \delta(m,c) + \delta(m,d) > \delta(c,d)$; this leads to a contradiction because (c,d) is a diameter pair.

- Prove that $mc = md$. Suppose $\delta(m,c) > \delta(m,d)$.

  We have : $2\delta(m,a) = 2\delta(m,b) = \delta(a,b) = \delta(c,d) \leq \delta(m,c) + \delta(m,d) < 2\delta(m,c)$. Therefore: $\delta(m,a) < \delta(m,c)$ and $\delta(m,b) < \delta(m,c)$.

  Case 1: c belongs to a different side of a to m. Then, $\delta(m,a) + \delta(m,c) = \delta(a,c) \implies \delta(m,a) + \delta(m,b) < \delta(a,c) \implies \delta(a,b) < \delta(a,c)$. This is a contradiction because (a,b) is a diameter pair of t.

  Case 2: c belongs to the same side of a to m. Then c belongs to a different side of b to m. Similar to case 1, in this case we can prove that $\delta(a,b) < \delta(b,c)$ which also leads to a contradiction.

Thus, m is the midpoint of the path between c and d.

□

This lemma allows us to define some new concepts that are useful in the rest of the proof.

New definitions:

The single midpoint of any tree t partitions the diameter set into disjoint subsets; we call each of those subsets a *diameter group* of t (if the midpoint is in the middle of the branch, we have two diameter groups; a midpoint coinciding on an internal node would give three or more groups). We call any restriction of t with k leaves removed a *k-optimal restricted tree* if no other restriction removing k leaves has a lower diameter. We call a tree t *k-shrinkable* if there exists a k-removing set that *strictly* reduces its diameter. We call any induced tree on t that has a smaller diameter than t a *shrunk tree* of t. Note that unless all but one of the diameter groups of a tree t are removed, the tree cannot shrink in diameter. When all but one of the diameter groups of a tree t is removed, we refer to the resulting tree as a *minimum shrunk tree* of t.

It is easy to see the following lemma.

Lemma 3. *For all a and b, $(a,b) \in \mathcal{P}(t)$ if and only if a and b belong to two distinct diameter groups.*

Now we prove a less obvious Lemma.

Lemma 4. *If tree t is k-shrinkable, any k-optimal restricted tree t^* can be induced from one of the minimum shrunk trees of t.*

Proof. Because t is k-shrinkable, the diameter of t^* must be strictly smaller than the diameter of t. Suppose t^* is not an induced tree of any minimum shrink tree of t; then, t* has at least two leaves from two different diameter groups of t. Based on Lemma 3, t^* shares with t at least one diameter pair and therefore, has the same diameter as t, which is a contradiction. □

We now turn to the proof of Theorem 3. Recall:

Theorem 3. For any k, any arbitrary pair-restricted k-removing space includes at least one optimal k-removing set.

Proof. If t is not k-shrinkable, any k-removing set is optimal and the result trivially follows. We now focus on a case where t is k-shrinkable.

Suppose t has m diameter groups:

$$D^1 = \{d_1^1, d_2^1, \dots, d_{p_1}^1\}, D^2 = \{d_1^2, d_2^2, \dots, d_{p_2}^2\}, \dots, D^m = \{d_1^m, d_2^m, \dots, d_{p_m}^m\} .$$

For $i = 1 \dots m$, let $k^i = |\bigcup_{j \neq i} D_j|$ be the size of all groups except group i, and let t^i denote the minimum shrunk tree of t that excludes all groups $D^j, j \neq i$. Let $k^p = \max_i(k^i)$. For the tree t to be k-shrinkable, we need that $k^i \leq k$; thus, $k \geq k^p$.

To produce any minimum shrunk tree t^i with $k^i \leq k^p$, we can start from any removal (a, b) such that $a \in D^x$ and $b \in D^y$ (for $x \neq y$), and continue to produce t^i. To see this, note that if $x \neq y \neq i$, any chain that starts with either a or b and continues to select from any groups other than D^i will produce the minimum shrunk tree t^i after k^i removals. Now, w.l.o.g, consider $x = i$ and $y \neq i$. Then, consider the chain that starts by removing y and continues by removals from any group other than D^i. This chain will also produce t^i after k^i removals. In other words, each pair-restricted k-removing space of t can produce all the minimum shrunk trees t^i that have $k^i \leq k$.

Based on Lemma 4, when t is k-shrinkable, at least one of the minimum shrunk trees (say t_i^*) can induce any k-optimal restricted tree t^*. We also just proved that any pair-restricted space can produce *all* minimum shrunk trees. Therefore, any arbitrary pair-restricted removing space will include a chain that induces t_i^* from t and another chain that produces t^* starting from t_i^*. Thus, the union of the removing sets corresponding to these two chains will produce t^* and will be part of any arbitrary pair-restricted k-removing space. □

A.2 Supplementary figures and tables

Sequence ID	Subtype
KJ723095	CRF01_AE
KJ723070	CRF02_AG
KJ723094	CRF02_AG
KJ723062	C
KJ723387	C
KJ723455	C
KJ723366	G
KJ722966	unassigned (B or CRF01_AE)
KJ723048	unassigned (B or F1)

Table A.1: Summary of the 9 outliers of the HIV dataset

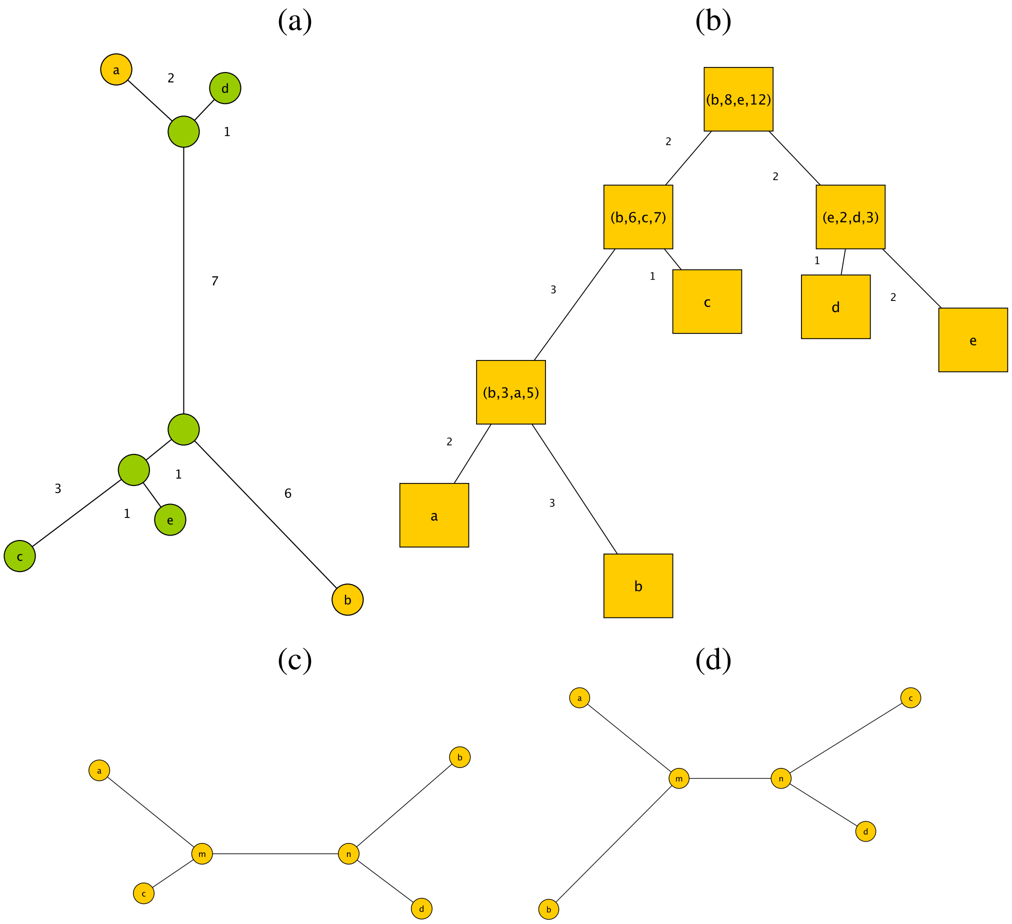


Figure A.1: Demonstration of the k-shrink algorithm. (*a*) An example where a greedy method does not work. This tree only has one diameter pair (a, b) (colored in yellow). If $k = 2$, the greedy method removes b and c and gives a restricted tree with diameter 11, while the optimal solution is to remove a and d, which gives the restricted tree with diameter 10. (*b*) The preprocessing step. In a post-order traversal, we store $rec(u)$ for each internal node u. The record has four values: leaf x under u that has the longest distance to u, the distance $\delta(u, x)$, the leaf y in one of the other sides of u to x that has the largest $\delta(x, y)$, and the distance. (c) and (d) Example quartet trees used in Proof of Lemma 1.

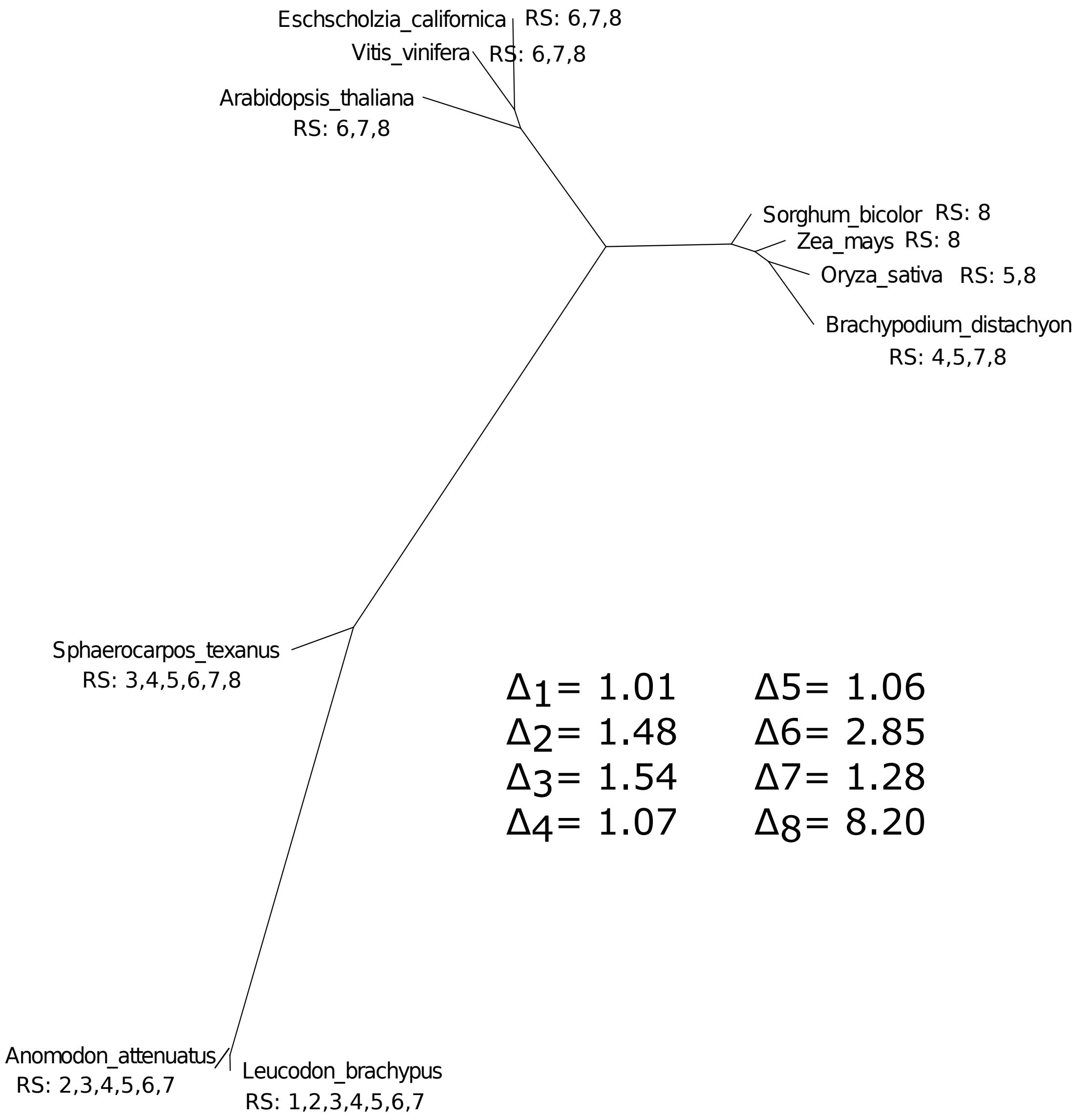
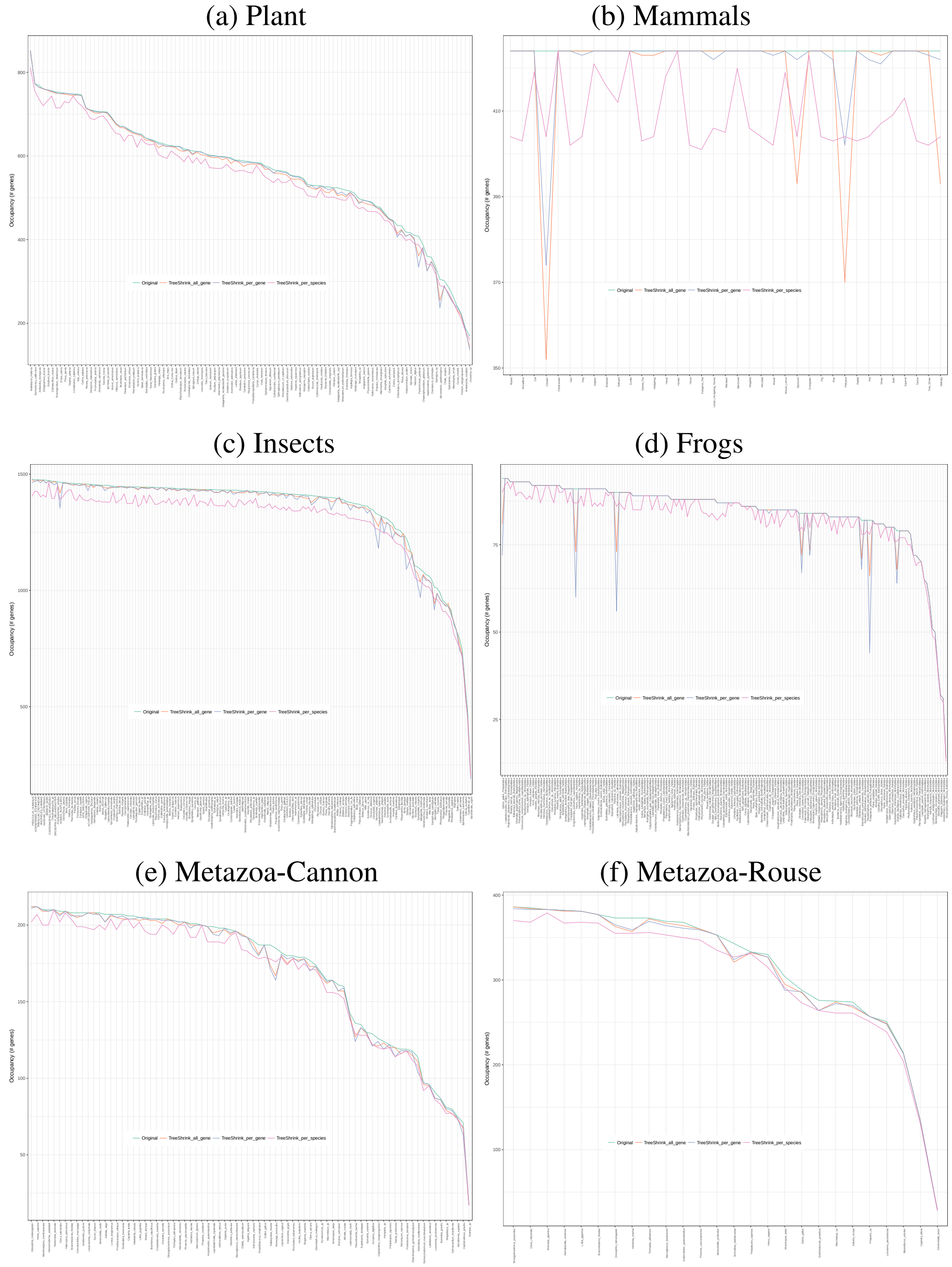

Eschscholzia_californica RS: 6,7,8
Vitis_vinifera RS: 6,7,8
Arabidopsis_thaliana RS: 6,7,8
Sorghum_bicolor RS: 8
Zea_mays RS: 8
Oryza_sativa RS: 5,8
Brachypodium_distachyon RS: 4,5,7,8
Sphaerocarpos_texanus RS: 3,4,5,6,7,8
Δ_1= 1.01
Δ_2= 1.48
Δ_3= 1.54
Δ_4= 1.07
Δ_5= 1.06
Δ_6= 2.85
Δ_7= 1.28
Δ_8= 8.20
Anomodon_attenuatus RS: 2,3,4,5,6,7
Leucodon_brachypus RS: 1,2,3,4,5,6,7

Figure A.2: An example of a very small tree from the Plants dataset. Solving *k*-shrink with $k = 8$ gives the removing set of all but two species Anomodon attenuatus and Leucodon brachypus. Although this removing set is obviously wrong, it has a very high ratio value ($\nu_8 = 8.197$).


(a) Plant

(b) Mammals

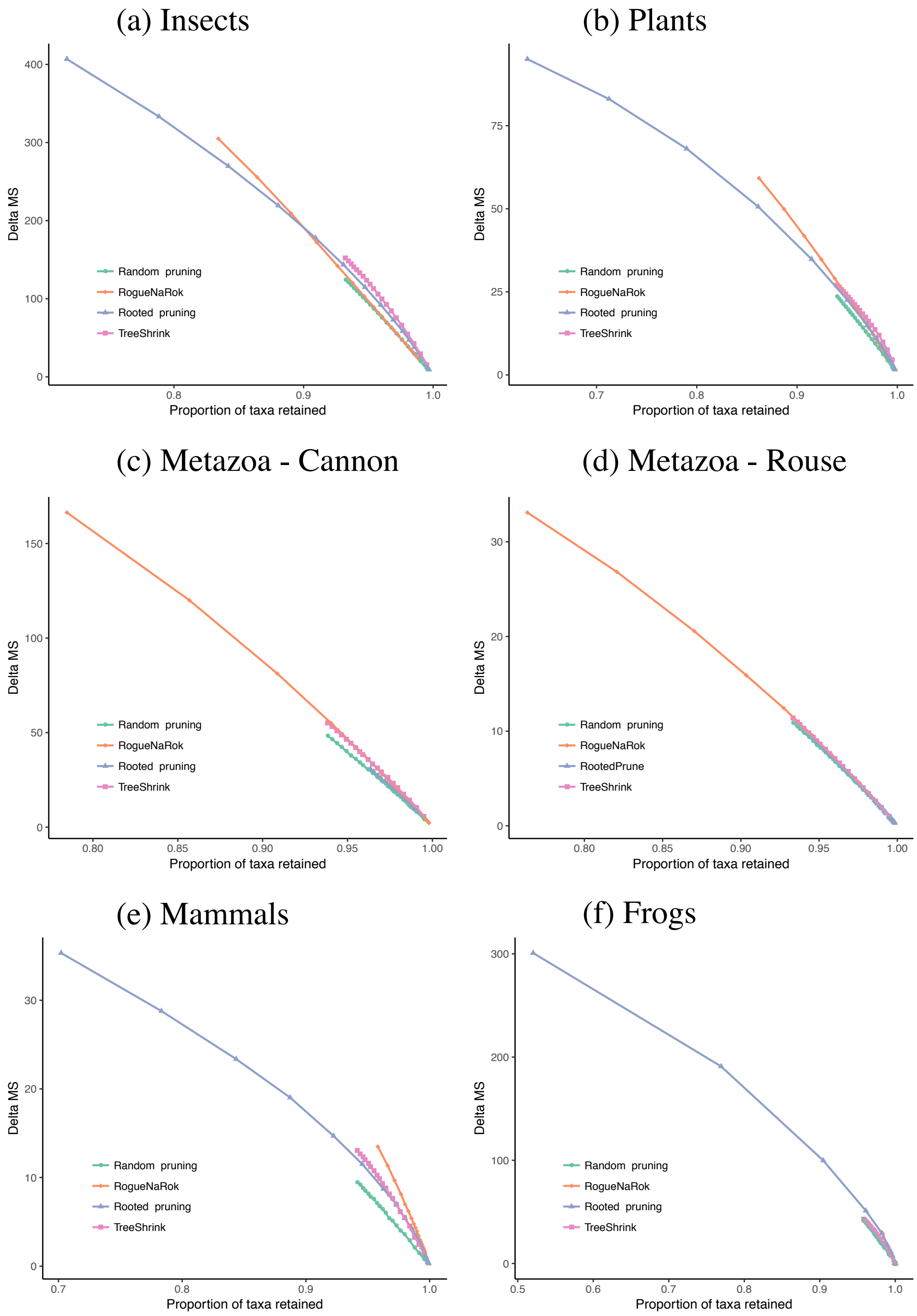
(c) Insects

(d) Frogs

(e) Metazoa-Cannon

(f) Metazoa-Rouse

Figure A.3: The impact of the 3 tests of TreeShrink on taxon occupancy for the six datasets. For each species (x-axis, ordered by occupancy), we show the number of genes that include it before and after filtering (default settings).



(a) Insects

(b) Plants

(c) Metazoa - Cannon

(d) Metazoa - Rouse

(e) Mammals

(f) Frogs

Figure A.4: The impact of TreeShrink, RogueNaRok, and rooted pruning on gene tree discordance on six datasets comparing to random pruning. MS distances are computed for all pairs of gene trees. The average reduction in the MS distance (*y-axis*) is shown versus the total proportion of the species retained in the gene trees after filtering (x-axis). A line is drawn between all points corresponding to different thresholds of the same method.

Appendix B

Supplementary materials for Chapter 2

B.1 Proof of Claim 1

Proof. First, note that any placement of x onto T results in a tree T' that shares all quartets with T except for the quartets that contain x. Therefore, to solve MQP, we only need to maximize the number of quartets that contain x and match R. As such, the MQP problem can be restated as: insert x to T to maximize the total number of quartets $xy|z_1z_2$ in T that match the corresponding quartet in R, where x, y, z_1, z_2 are distinct and $y, z_1, z_2 \in L_T \cap L_R$. Consider an arbitrary rooting of T. Observe that if we add x as an outgroup to T to obtain T_x, then there is a one-to-one mapping of every quartet $xy|z_1z_2$ in T_x to the triplet $y|z_1z_2$ in T. Similarly, each quartet $xy|z_1z_2$ of R has a unique mapping to the triplet $y|z_1z_2$ in R' (recall that R' has been rooted at x before having it removed). Therefore, if $y|z_1z_2$ of T matches R', then $xy|z_1z_2$ of T_x matches R. Thus, to find T_x that has the maximum number of matching quartets to R, we solve MTR on T and R' to maximize the number of matching triplets $y|z_1z_2$, then add x as an outgroup and obtain the maximum number of matching quartets $xy|z_1z_2$. □

B.2 Proof of the $\pi_j^{C_1C_2\to C}$ equation

$\pi_j^{C_1C_2\to C}$ is the number of triplets in the following set:

$$\{(ii|k)\in C|i,k\neq j,i\neq k\}\cup\{(ikm)\in C|i\neq k\neq m\neq j\}, \tag{B.1}$$

where i,j,k,m denotes a leaf color and $(ab|c)$ denotes the resolved triplet that separates (a,b) from c and (abc) denotes an unresolved triplet on $\{a,b,c\}$. The number of triplets in this set is the sum of $\pi_j^{C_1}, \pi_j^{C_2}$, and the following 4 cases (see Fig B.1a for an illustration):

Case 1: The number of ways to choose one i-colored leaf in C_1 (i.e. $n_i^{C_1}$) and one pair of leaves with colors i and k in C_2 such that the i-colored leaf is below the k-colored leaf in C_2 (i.e. $n_{i\uparrow k}^{C_2}$), for all choices of i,k such that i,j,k are pairwise distinct. Thus, the counter in case 1 is

$$\sum_{i\neq j} n_i^{C_1} \sum_{k\neq i,k\neq j} n_{i\uparrow k}^{C_2} = \sum_{i\neq j} n_i^{C_1}(n_{i\uparrow\bullet}^{C_2} - n_{i\uparrow j}^{C_2}) \tag{B.2}$$

Case 2: The number of ways to choose two i-colored leaves in C_1 (i.e. $\binom{n_i^{C_1}}{2}$) and one k-colored leaf in C_2 (i.e. $n_k^{C_2}$), for all choices of i,k such that i,j,k are pairwise distinct. Thus, the counter in case 2 is

$$\sum_{i\neq j}\binom{n_i^{C_1}}{2}\sum_{k\neq i,k\neq j}(n_k^{C_2}) = \sum_{i\neq j}\binom{n_i^{C_1}}{2}(n_\bullet^{C_2} - n_j^{C_2} - n_i^{C_2}) \tag{B.3}$$

Case 3: The number of ways to choose one k-colored leaf in C_1 (i.e. $n_k^{C_1}$) and two i-colored leaves in C_2 such that the LCA of these two i-colored leaves is NOT on the external path linking C_1 and C_2 (i.e. $n_{(i,i)}^{C_2}$), for all choices of i,k such that i,j,k are pairwise distinct. Thus, the counter in case 3 is

$$\sum_{i\neq j,k\neq j|i\neq k} n_k^{C_1} n_{(ii)}^{C_2} = \sum_{i\neq j}(n_{\bullet}^{C_1} - n_j^{C_1} - n_i^{C_1}) n_{(ii)}^{C_2} \tag{B.4}$$

Case 4 (unresolved triplets): The number of ways to choose one *i*-colored leaf in C_1 (i.e. $n_i^{C_1}$), and one *k*-colored leaf in C_2 and one *m*-colored leaf in C_2 such that their LCA is on the external path linking C_1 and C_2 and furthermore, the path from their LCA to either the *k*-colored or *m*-colored leaf does not consist any other node on the external path (i.e. $n_{km}^{C_2}$), for all choices of i,k,m such that i,j,k,m are pairwise distinct. Thus, the counter in case 4 is

$$\sum_{i,k,m|i,j,k,m \texttt{ are pairwise distinct}} n_i^{C_1} n_{km}^{C_2} = \sum_{i\neq j} n_i^{C_1}(n_{\bullet\square}^{C_2} - n_{j\bullet}^{C_2} - n_{i\bullet}^{C_2} + n_{ij}^{C_2}) \tag{B.5}$$

Summing up all the four cases and adding $\pi_j^{C_1}$, $\pi_j^{C_2}$, we get the equation as in the main text.

B.3 Proof of the $\rho^{C_1C_2\to C}$ equation

$\rho^{C_1C_2\to C}$ is the number of triplets in the following set:

$$\{(00|k)\in C|k\neq 0\}\cup\{(ij|0)\in C|i\neq 0, j\neq 0\}, \tag{B.6}$$

where i,j,k denotes a leaf color and $(ab|c)$ denotes the resolved triplet that separates (a,b) from c and (abc) denotes an unresolved triplet on $\{a,b,c\}$. The number of triplets in this set is the sum of ρ^{C_1}, ρ^{C_2}, and the triplets in one of the three cases that are similar to cases 1, 2, and 3 of $\pi_j^{C_1C_2\to C}$. For each of these 3 cases, we can convert the equation to compute $\pi_j^{C_1C_2\to C}$ to $\rho^{C_1C_2\to C}$ by summing up the following two counters (see Fig. B.1b for an illustration of the two counters in case 1).

- Replace the two *i*-colored leaves with two 0-colored ones. Each of these triplets has the

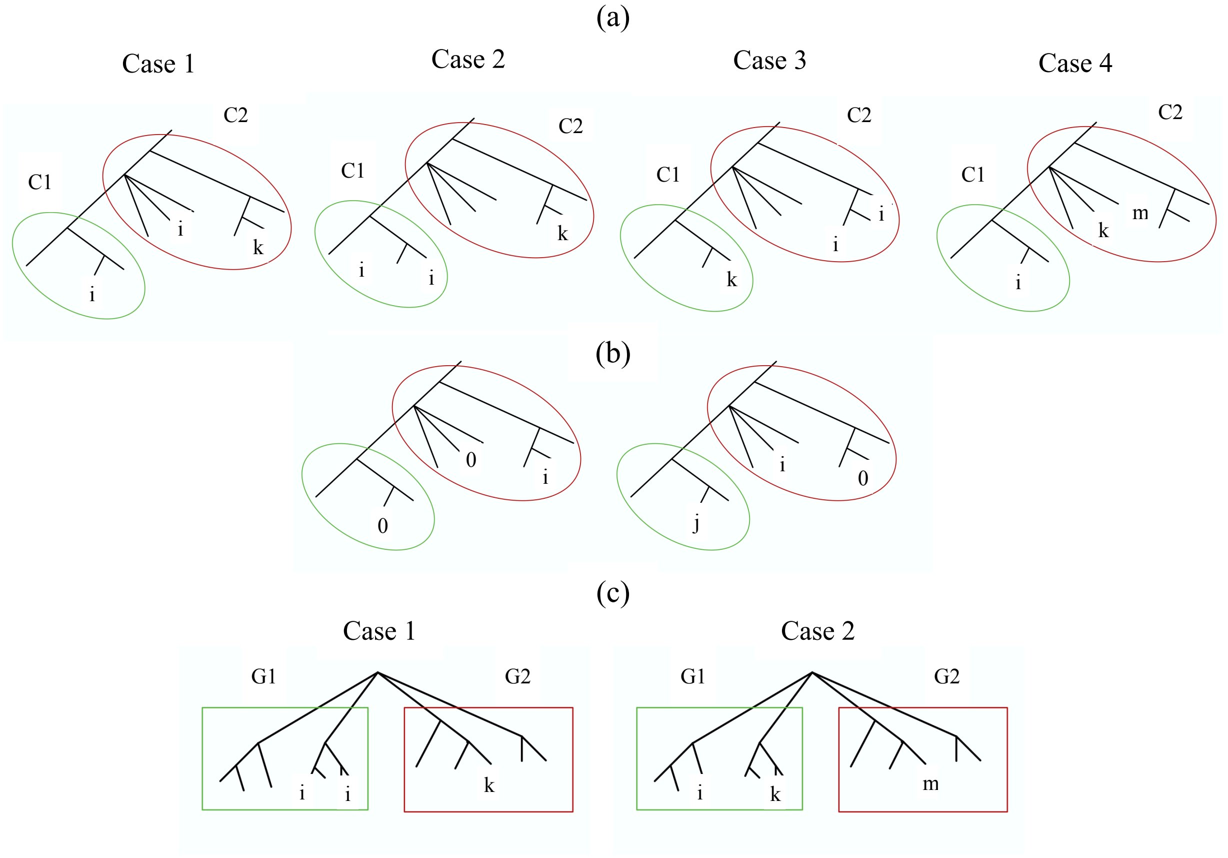


Figure B.1: (a) The 4 cases of triplet arrangement for $\pi_j^{C_1C_2 \to C}$. In each case, we assume i, k, m are distinct and different from j. (b) The two counters for $\rho^{C_1C_2 \to C}$ in case 1. Here we assume i and j are different from 0, but are not necessarily distinct. (c) The 2 cases of triplet arrangement for $\pi_j^{G_1G_2 \to G}$. In each case, we assume i, k, m are distinct and different from j. To compute the complete counter, we also need to swap the roles of G_1 and G_2 to count all possible triplets.

form $(00|k)$. To get the counter, we count all the different ways to choose the two 0-colored leaves and the k-colored leaf over all possible choices of k.

- Use a 0-colored leaf in place of the k-colored and use a j-colored leaf (for any $j \neq 0$, possibly equal to i) in place of one of the two i-colored leaves. It is easy to see that each of these triplets (in any of the cases 1, 2, 3) has the form $(ij|0)$. To get the counter, we count all the different ways to choose the 0-colored i-colored, and j-colored leaves over all possible choices of i and j.

It is straightforward to derive the equation of each counter for each case 1, 2, 3. Summing up all the counters across all cases then adding ρ^{C_1} and ρ^{C_2}, we get the equation as in the main text.

B.4 Proof of the $\pi_j^{G_1 G_2 \to G}$ equation

$\pi_j^{G_1 G_2 \to G}$ is the number of triplets in the following set:

$$\{(ii|k) \in G | i,k \neq j, i \neq k\} \cup \{(ikm) \in G | i \neq k \neq m \neq j\}, \tag{B.7}$$

where i, j, k, m denotes a leaf color and $(ab|c)$ denotes the resolved triplet that separates (a,b) from c and (abc) denotes an unresolved triplet on $\{a,b,c\}$. The number of triplets in this set is the sum of $\pi_j^{G_1}, \pi_j^{G_2}$, and the following 2 cases (see Fig. B.1c for an illustration):

Case 1: The number of ways to choose one k-colored leaf in G_2 (i.e. $n_k^{G_2}$) and two i-colored leaves in G_1 such that the LCA of these two i-colored leaves is NOT the super root of G (i.e. $n_{(i,i)}^{G_1}$), for all choices of i,k such that i, j, k are pairwise distinct. Additionally, we also swap the role of G_1 and G_2 and add to the counter. Thus, the counter in case 1 is

$$\sum_{i\neq j,k\neq j|i\neq k} n_k^{G_2} n_{(ii)}^{G_1} + \sum_{i\neq j,k\neq j|i\neq k} n_k^{G_1} n_{(ii)}^{G_2} = \sum_{i\neq j} (n_\bullet^{G_2} - n_j^{G_2} - n_i^{G_2}) n_{(ii)}^{G_1} + \sum_{i\neq j} (n_\bullet^{G_1} - n_j^{G_1} - n_i^{G_1}) n_{(ii)}^{G_2} \tag{B.8}$$

Case 2 (unresolved triplets): The number of ways to choose one i-colored leaf in G_1 (i.e. $n_i^{G_1}$), and one k-colored and one m-colored leaves in G_2 such that their LCA is the super root of G (i.e. $n_{km}^{G_2}$), for all choices of i,k,m such that i,j,k,m are pairwise distinct. Additionally, we also swap the role of G_1 and G_2 and add to the counter. Thus, the counter in case 2 is

$$\begin{aligned} &\textstyle\sum_{i,k,m|i,j,k,m \texttt{ are pairwise distinct}} (n_i^{G_1} n_{km}^{G_2} + n_i^{G_2} n_{km}^{G_1}) \\ &= \textstyle\sum_{i\neq j} n_i^{G_1} (n_{\bullet\square}^{G_2} - n_{j\bullet}^{G_2} - n_{i\bullet}^{G_2} + n_{ij}^{G_2}) + \sum_{i\neq j} n_i^{G_1} (n_{\bullet\square}^{G_2} - n_{j\bullet}^{G_2} - n_{i\bullet}^{G_2} + n_{ij}^{G_2}) \end{aligned} \tag{B.9}$$

Summing up the two cases and adding $\pi_j^{G_1}$, $\pi_j^{G_2}$, we get the equation as in the main text.

B.5 Proof of the $\rho^{G_1 G_2 \to G}$ equation

$\rho^{G_1 G_2 \to G}$ is the number of triplets in the following set:

$$\{(00|k) \in G | k \neq 0\} \cup \{(ij|0) \in G | i \neq 0, j \neq 0\}, \tag{B.10}$$

where i,j,k denotes a leaf color and $(ab|c)$ denotes the resolved triplet that separates (a,b) from c and (abc) denotes an unresolved triplet on $\{a,b,c\}$. The number of triplets in this set is the sum of ρ^{G_1}, ρ^{G_2}, and the triplets in the case corresponding to case 1 of $\pi_j^{G_1 G_2 \to G}$. Following the same reasoning as in $\rho^{C_1 C_2 \to C}$, we can easily derive the equation for $\rho^{G_1 G_2 \to G}$ from $\pi_j^{G_1 G_2 \to G}$. Summing up this counter and ρ^{G_1}, ρ^{G_1}, we get the equation as in the main text.

Type	Description
L	derived from a leaf in R. Any L node is a leaf of the HDT and is linked to the corresponding leaf in T
I	a node of the HDT derived from an internal node in R. Any I node is a *leaf* of the HDT
G	derived from a set of subtrees in R with roots being siblings. A G node can either be a L or $GG \to G$
C	derived from a connected subset of nodes in R. A C node can either be a $IG \to C$ or $CC \to C$
$GG \to G$	a node of the HDT that is the parent of two other G nodes
$IG \to C$	a node of the HDT that is the parent of an I and a G node
$CC \to C$	a node of the HDT that is the parent of two C nodes

Table B.1: HDT node types. The G and C are super-types.

B.6 Supplementary figures

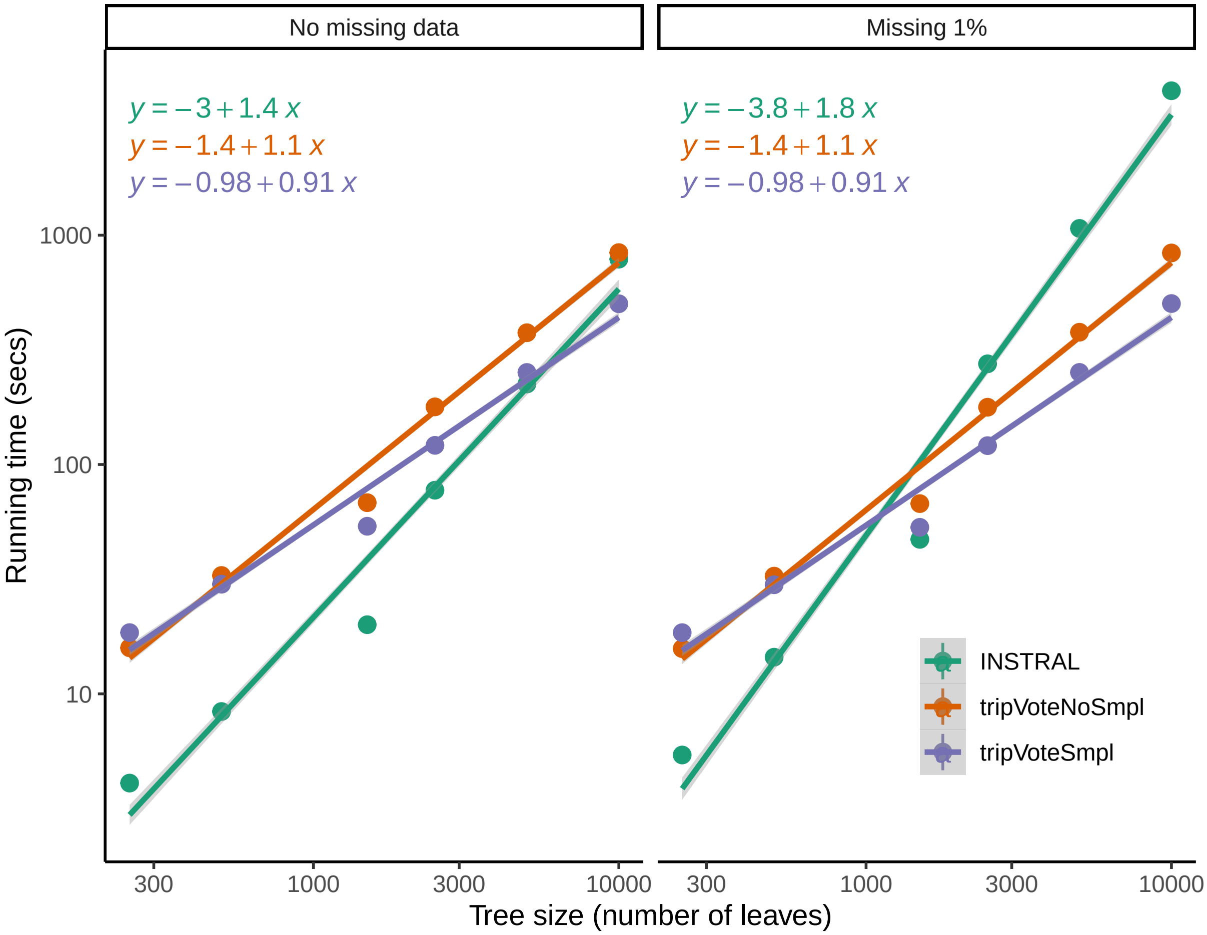


Figure B.2: Run time of tripVote and INSTRAL (seconds) on different trees with sizes. Left panel: no missing data and Right panel: missing data is introduced by removing one percent of the taxa from each gene tree.

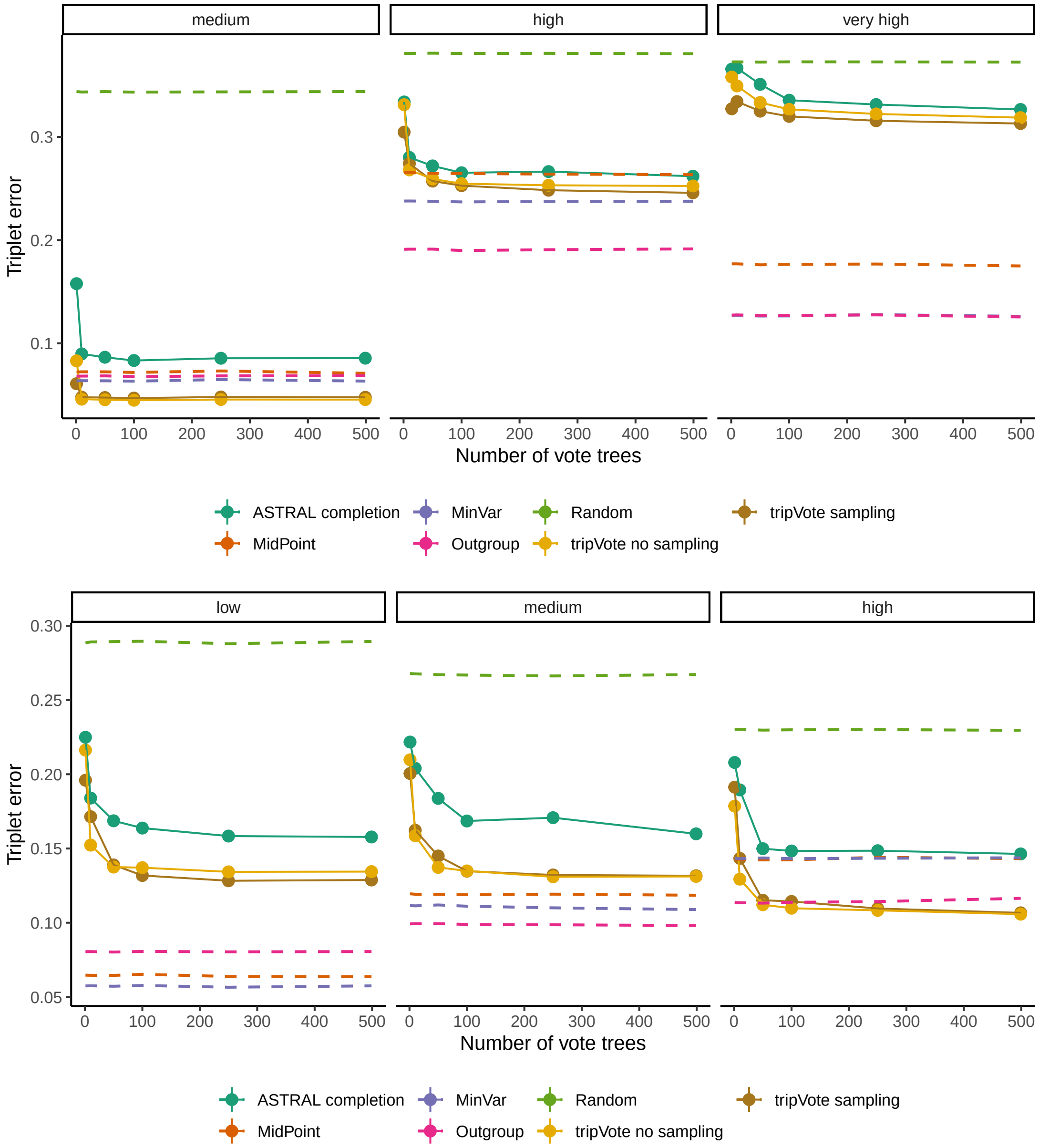


Figure B.3: Accuracy of rooting based on different methods for Top: The 201-taxon dataset and Bottom: The 31-taxon dataset. The outgroup is removed from *m* randomly selected trees and inserted back using either ASTRAL completion or tripVote, then each of these trees is rerooted at the reinserted outgroup. The x-axis shows the number of voting trees for ASTRAL completion and tripVote (i.e. $n-m$) and the y-axis shows the triplet error to the true rooted tree. We added alternative rooting methods (Outgroup, MinVar, MidPoint, and Random) that do not use other gene trees. Outgroup rooting was run on the complete estimated trees *before* removing the outgroup. MidPoint and MinVar were run *after* the outgroup is removed. The Random rooting was repeated 50 times and the reported error is the average.

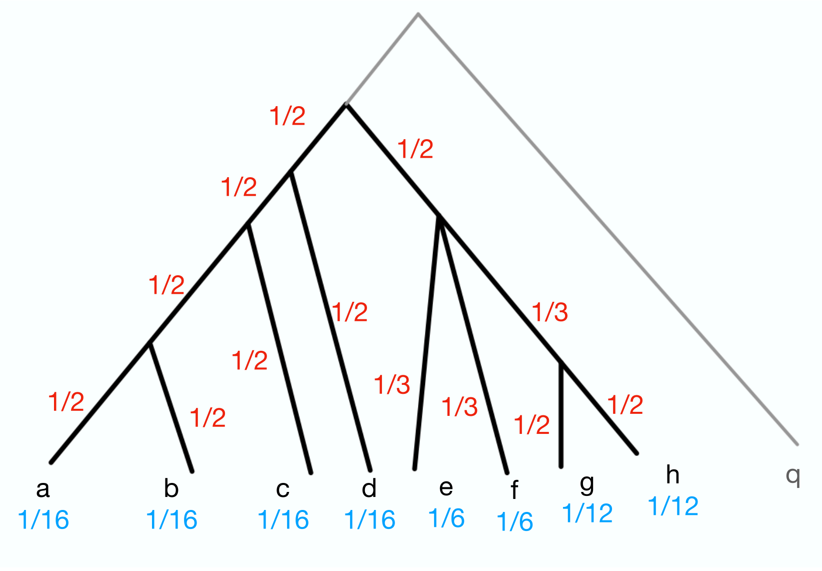


Figure B.4: Given a reference tree obtained by rooting a tree on a query leaf q and removing q, we sample as follows. We follow a random path from the root to a leaf, at each internal node u, selecting a child uniformly at random; the leaf node is sampled. Equivalently, we sample from a multinomial distribution with sampling probability equal to the product of the probabilities of the edges along the path from the root to that leaf.

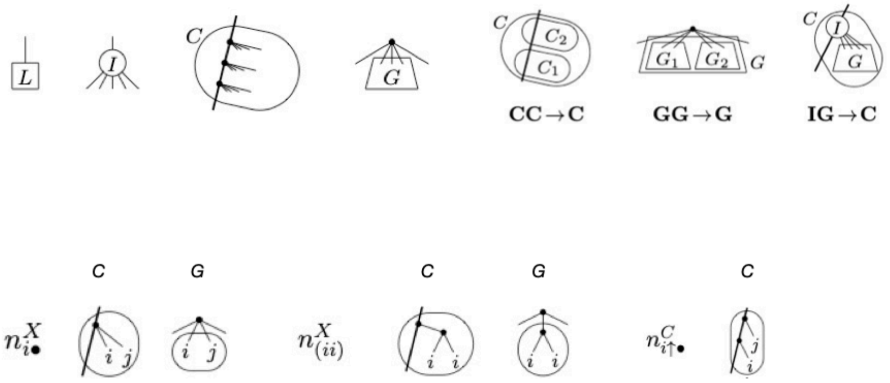


Figure B.5: Top: Different types of components in the HDT. (figure taken from [153]). Bottom: Different counters maintained at each component C or G in the HDT (figure taken from [153]).

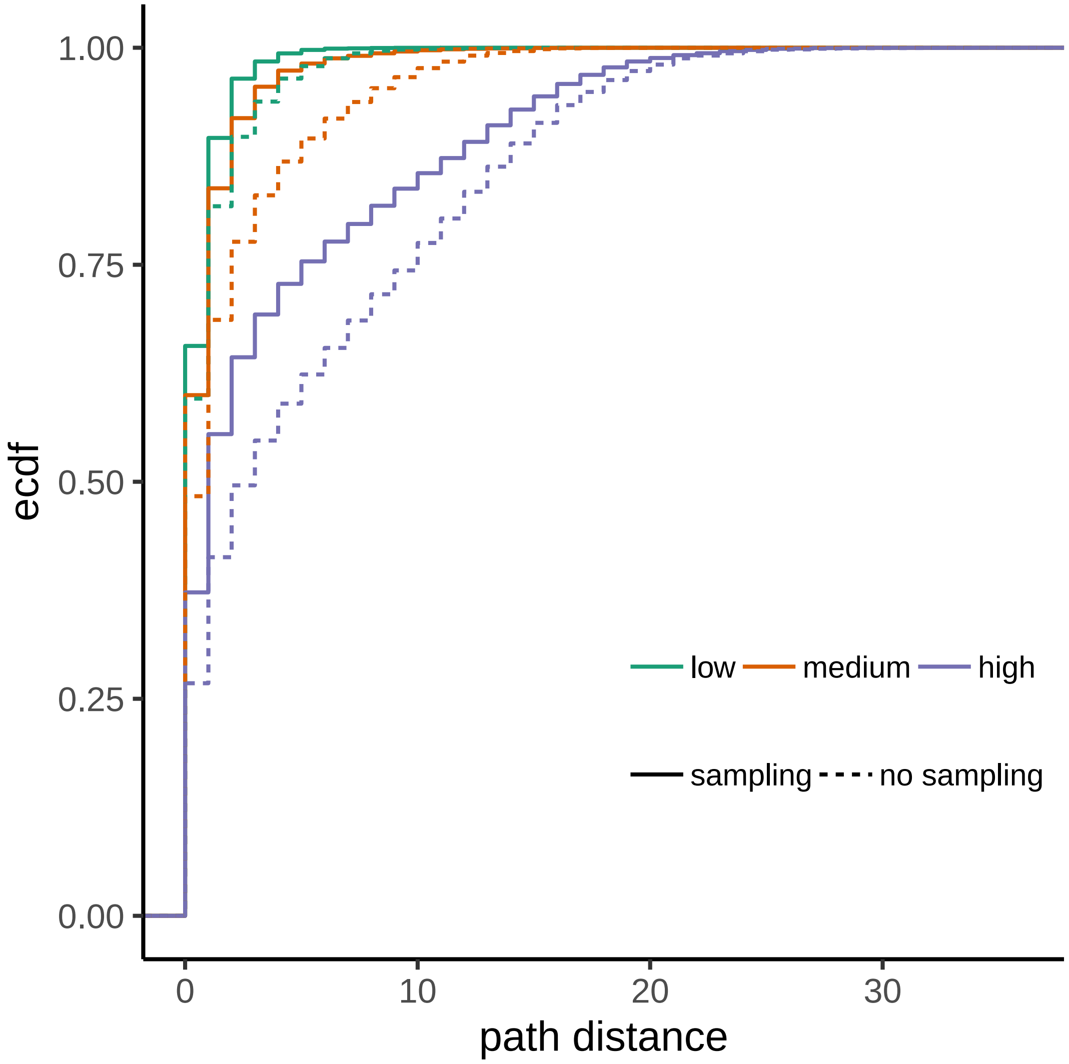


Figure B.6: ECDF plot of placement error by tripVote with and without sampling strategy on different levels of ILS. In this experiment, *m* is set to 1 and the error is measured by the path distance between the true placement branch (from the full estimated tree by FastTree) and the estimated placement branch by tripVote.

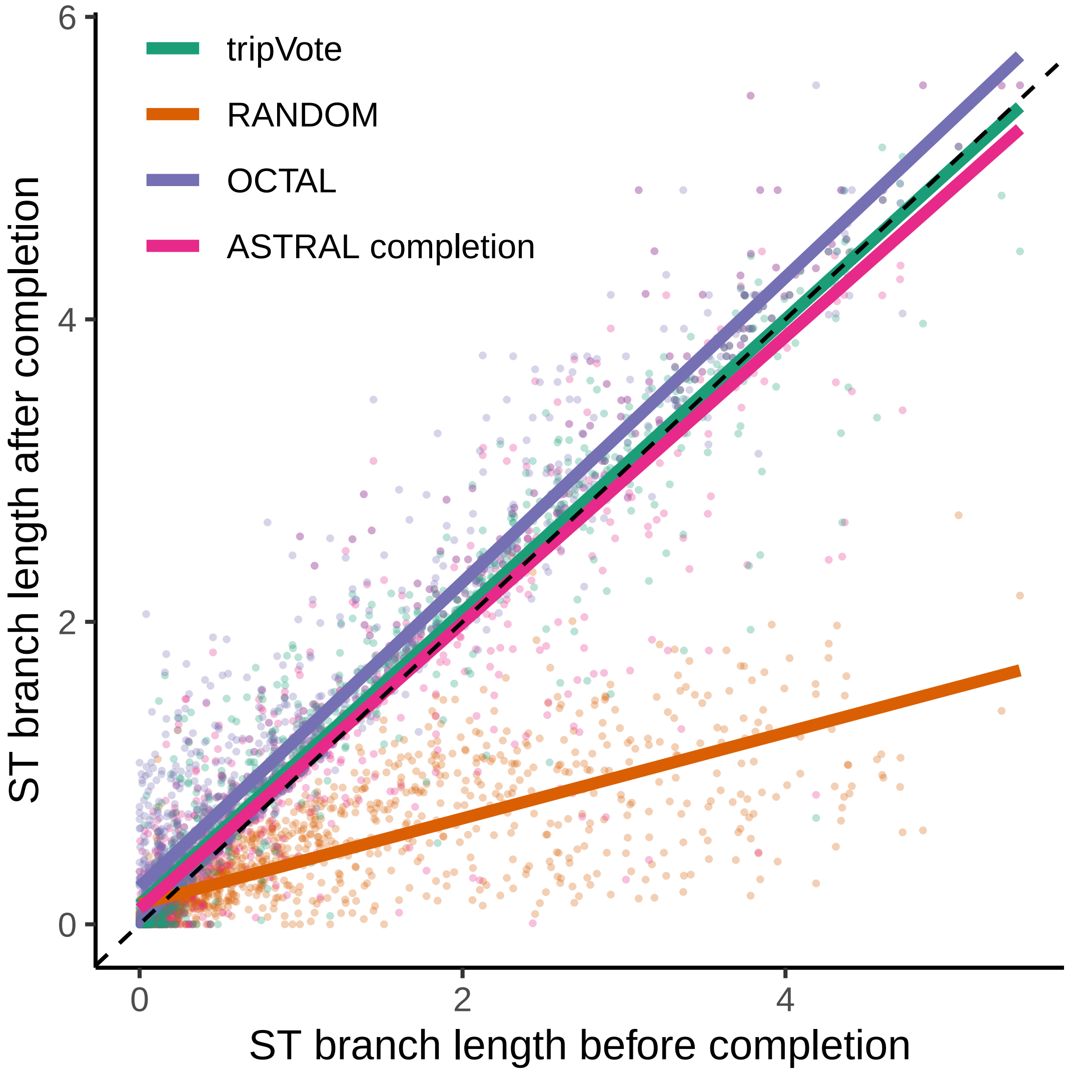


Figure B.7: The estimated branch lengths of the species tree based on the original incomplete gene trees versus the completed gene trees using different methods. In all scenarios, the species tree topology is fixed and the branch lengths are estimated using ASTRAL-III.

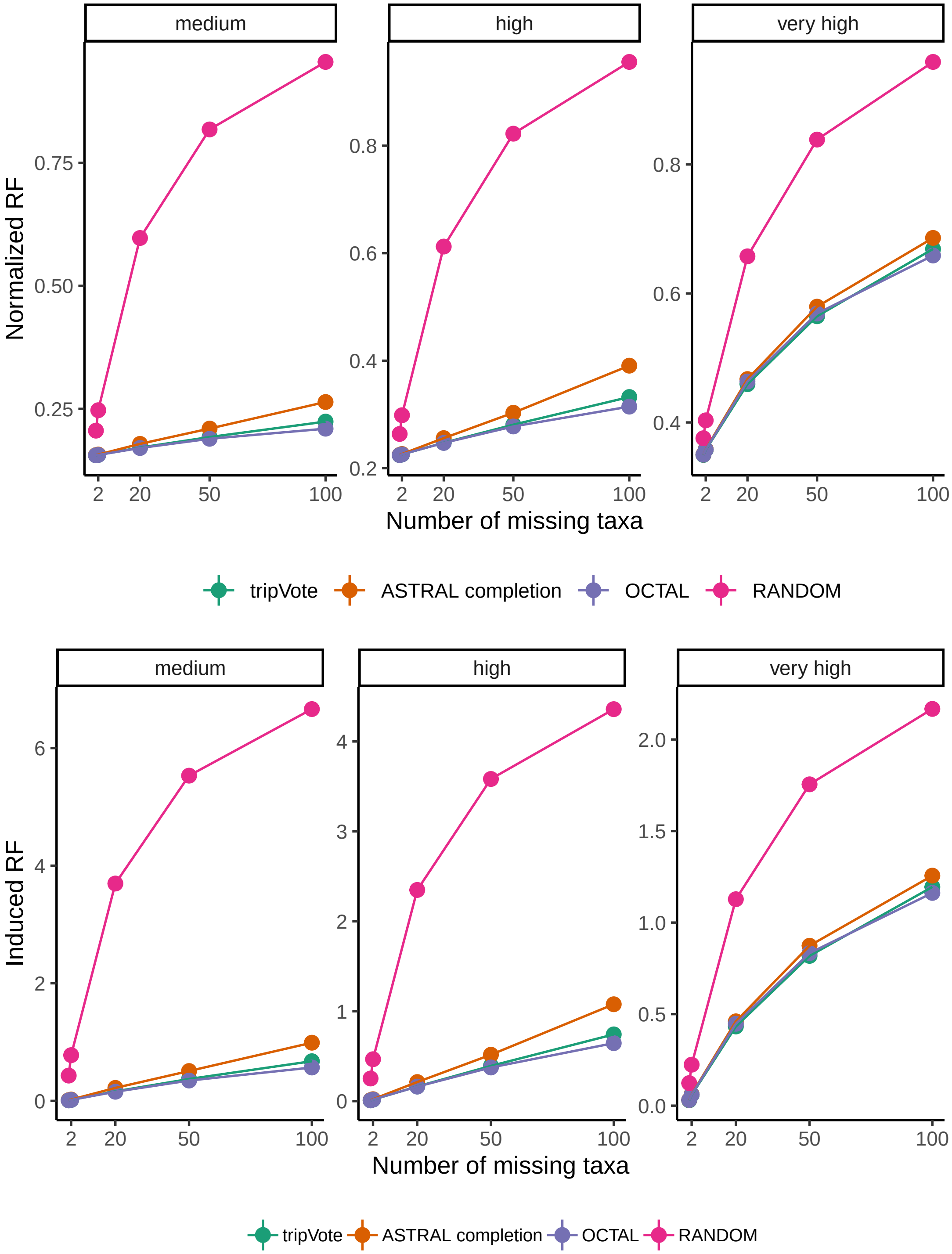


Figure B.8: Accuracy of different methods on completing gene trees of the 201-taxon dataset. Top: Normalized RF error and Bottom: Induced RF error.

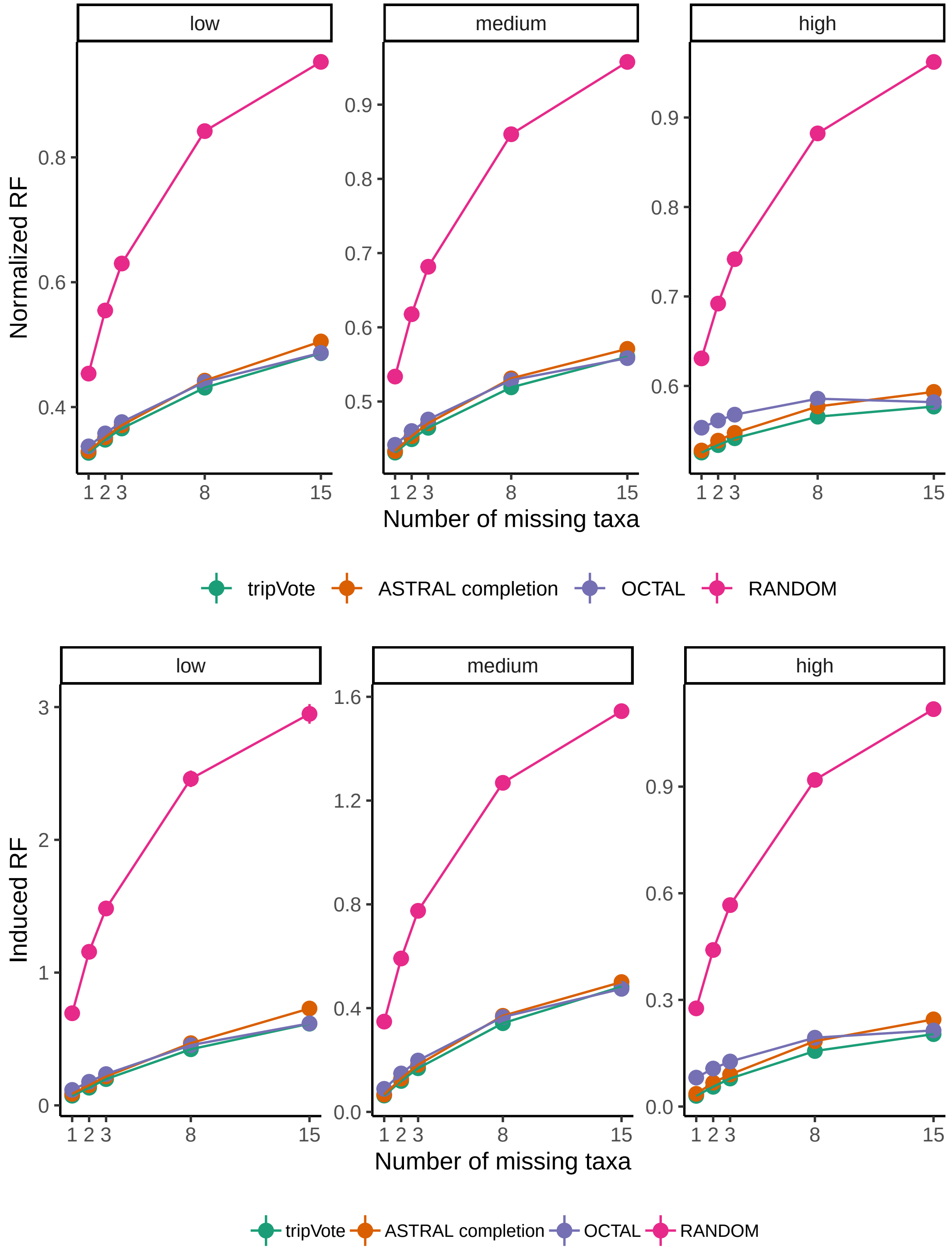


Figure B.9: Accuracy of different methods on completing gene trees of the 31-taxon dataset. Top: Normalized RF error and Bottom: Induced RF error.

Appendix C

Supplementary materials for Chapter 3

C.1 Supplementary theory

In this section we prove the following propositions, which were used in the main text to support the theory of the MinVar rooting method. Please refer to the main paper for more details.

Proposition 3: *A point p on tree T is a local MV if and only if it is a balance point.*

Based on Proposition 3, we refer to local MV and balance point interchangeably.

Proposition 4: *Any tree has at least one local MV.*

Proposition 5: *The global MV of any tree is one of its local MVs.*

Proposition 6: *Let p denote the global MV of T. If*

$$\varepsilon \leq \min_{w \in c(r)} \left(\frac{e_w}{\frac{n}{n-|w|} h + e_w} \right)$$

then there exists a child w of r such that $p \in e(r, w)$.

Proposition 7: *When the global MV is on one of the adjacent edges of r, let a random variable X indicate the distance of the global MV to the root; then, $E(X) = 0$.*

Proposition 8: *Let p be a point on an edge (u, v) of tree T with distance $d(p, u) = x$. If we let p vary along edge (u, v) and consider $var(p)$ as a function of variable x with parameters u and v,* *then:*

$$var(p) = var(x; u, v) = (1-\beta^2)x^2 + \left(\alpha - \frac{2ST(u)\beta}{n}\right)x + var(u) \tag{C.1}$$

in which

$$\alpha = \frac{2ST(u) - 4(SI(v) + |v|e_v)}{n} \quad \text{and} \quad \beta = 1 - \frac{2|v|}{n} \tag{C.2}$$

C.1.1 Extra notations

For two points p and p', potentially on different edges, we let $path(p, p')$ denote the *directed* path from p to p'. For two nodes p and u, we define $Cld_p(u)$ as the clade under u if the tree T is rerooted at p. For ease of notation we use $|p \triangleright u|$ to denote the size of $Cld_p(u)$. For a point p on tree T and another point p' on either the same edge or an edge connected to p (if p is a node), we let $\overrightarrow{pp'}$ denote a *direction* of p. It is easy to see that any point on a tree has at least two directions, and any node that is not the root has at least three directions. We call $\overrightarrow{pp'}$ a *dominant* direction of p if and only if

$$\frac{1}{|p \triangleright p'|} \sum_{i \in Cld_p(p')} d_i(p) > \frac{1}{n - |p \triangleright p'|} \sum_{i \notin Cld_p(p')} d_i(p) \tag{C.3}$$

C.1.2 Proofs

Proofs of ST relation and Proposition 8

On a tree T, consider a point p on the edge (u, v) with distance x from u (Fig C.1).

Proof of ST relation. Recall that $ST(v)$ is the sum of distances of all leaves from the node v (i.e. $ST(p) = \sum_{i \in Cld(p)}(d_i(p))$. We need to prove that

$$ST(v) = ST(p(v)) + (n - 2|v|)e_v. \tag{C.4}$$

We have

$$\begin{aligned} ST(p) &= \sum_{i \in Cld(p)} (d_i(u) - x) + \sum_{i \in L - Cld(p)} (d_i(u) + x) \\ &= \sum_{i \in L} d_i(u) + (|L| - |p| - |p|)x \qquad \text{(C.5)} \\ &= ST(u) + (n - 2|p|)x \end{aligned}$$

Let $p \equiv v$, we get Eq. C.4. □

Proof of Proposition 8. Recall that $ST(p) = \sum_{i \in L} d_i(p)$.

$$var(p) = \frac{1}{n}\sum_{i \in L}(d_i(p) - \frac{\sum_{i \in L} d_i(p)}{n})^2 = \frac{\sum_{i \in L} d_i^2(p)}{n} - (\frac{ST(p)}{n})^2 \qquad \text{(C.6)}$$

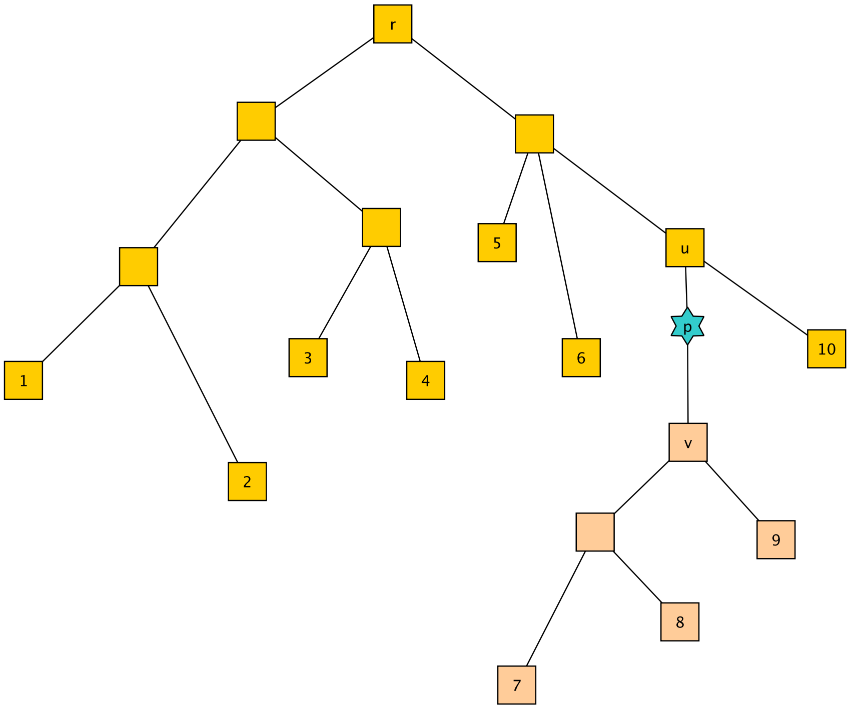


Figure C.1: An example tree T rooted at r with a point p on edge (u, v).

The first term of the RHS of C.6 can be expanded as follow:

$$\begin{aligned}
\frac{\sum_{i\in L} d_i^2(p)}{n} &= \frac{1}{n}\sum_{i\in Cld(v)}(d_i(u)-x)^2 + \frac{1}{n}\sum_{i\in L-Cld(v)}(d_i(u)+x)^2 \\
&= \frac{1}{n}\sum_{i\in Cld(v)}(d_i^2(u)-2d_i(u)x+x^2) + \frac{1}{n}\sum_{i\in L-Cld(v)}(d_i^2(u)+2d_i(u)x+x^2) \\
&= \frac{1}{n}\sum_{i\in L} d_i^2(u) + \frac{2}{n}(\sum_{i\in L-Cld(v)} d_i(u) - \sum_{i\in Cld(v)} d_i(u))x + x^2 \\
&= \frac{1}{n}\sum_{i\in L} d_i^2(u) + 2x\frac{\sum_{i\in L} d_i(u) - 2\sum_{i\in Cld(v)} d_i(u)}{n} + x^2 \qquad \text{(C.7)} \\
&= \frac{1}{n}\sum_{i\in L} d_i^2(u) + 2x\frac{ST(u) - 2\sum_{i\in Cld(v)}(d_i(v)+e_v)}{n} + x^2 \\
&= \frac{1}{n}\sum_{i\in L} d_i^2(u) + 2x(\frac{ST(u) - 2(SI(v)+|v|e_v)}{n}) + x^2 \\
&= \frac{1}{n}\sum_{i\in L} d_i^2(u) + \alpha x + x^2
\end{aligned}$$

where the last line is simply derived from the definition:

$$\alpha = \frac{2ST(u) - 4(SI(v)+|v|e_v)}{n}$$

Recall $\beta = (1 - \frac{2|v|}{n})$; the second term can be expanded as follow:

$$\begin{aligned}
\left(\frac{ST(p)}{n}\right)^2 &= \left(\frac{ST(u)+(n-2|v|)x}{n}\right)^2 \\
&= \left(\frac{ST(u)}{n} + \beta x\right)^2 \qquad \text{(C.8)} \\
&= \left(\frac{ST(u)}{n}\right)^2 + \frac{2ST(u)\beta x}{n} + \beta^2 x^2
\end{aligned}$$

Substitute C.7 and C.8 to C.6, we obtain:

$$
\begin{aligned}
var(p) &= \frac{\sum_{i\in L} d_i^2(u)}{n} + \alpha x + x^2 - \left(\frac{ST(u)}{n}\right)^2 - \frac{2ST(u)\beta x}{n} - \beta^2 x^2 \\
&= \frac{\sum_{i\in L} d_i^2(u)}{n} - \left(\frac{ST(u)}{n}\right)^2 + \left(\alpha - \frac{2ST(u)\beta}{n}\right)x + (1-\beta^2)x^2 \qquad \text{(C.9)} \\
&= var(u) + \left(\alpha - \frac{2ST(u)\beta}{n}\right)x + (1-\beta^2)x^2
\end{aligned}
$$

Thus, we get Eq. 3.4 □

Useful lemmas

Below are useful lemmas that will be used later in the proofs.

Lemma 5. *Any point on a tree either is a balance point or has at least one dominant direction.*

Proof. On tree T, consider an arbitrary point p that is adjacent to nodes $v_1, v_2, ..., v_k$ of T. Let $\mu_j = \frac{1}{|p \triangleright v_j|}\sum_{i\in Cld_p(v_j)} d_i(p)$. If $\mu_1 = \mu_2 = ... = \mu_k$, then p is a balance point of T. Otherwise, let $\mu_m = max(\mu_1, \mu_2, ..., \mu_k)$. It is easy to see that $\overrightarrow{pv_m}$ is a dominant direction of p. □

Lemma 6. *If a point p_0 is not a local MV of tree T, there exists at least one point p' on T such that $var(p') < var(p_0)$.*

Lemma 7. *Consider an edge $e = (u, v)$ of tree T. If $\overrightarrow{uv}$ is a dominant direction of u and $\overrightarrow{vu}$ is a dominant direction of v, then there exists a balance point on edge e.*

(Lemmas 6 and 7 are proved later)

Proofs of Proposition 3 and Lemma 6

We start by some definitions and derivations that are used in proofs of both Proposition 3 and Lemma 6. Consider a point p_0 on tree T and any arbitrary point p on the same edge as p_0 or on an edge adjacent to p_0 if p_0 is a node. Note that when p_0 is in the middle of a edge, p can be a point above or below it on the same edge, but when p_0 is a node, p can be a point on any of the three (or more) edges adjacent to p. We divide the leaf set L of T into two disjoint groups: the leaves inside $Cld_{p_0}(p)$ (group 1), and the remaining leaves (group 2). Let $x = d(p_0, p)$, n be the size of T, and k be the size of group 1; the size of group 2 is therefore $n-k$. Let $d'_1, d'_2, ..., d'_k$ be the distances of the leaves in group 1 to p_0, $d'_{k+1}, d'_{k+2}, ..., d'_n$ be the distances of the leaves in group 2 to p_0, $d_1, d_2, ..., d_k$ be the distances of the leaves in group 1 to p, and $d_{k+1}, d_{k+2}, ..., d_n$ be the distances of the leaves in group 2 to p. Also let μ' and μ be the averages of the leaf distances to p_0 and p. Then:

$$d_i = \begin{cases} d'_i - x, & \text{if } 1 \le i \le k \\ d'_i + x, & \text{if } k+1 \le i \le n \end{cases} \tag{C.10}$$

$$\mu' = \frac{1}{n}\left(\sum_{i=1}^{n} d'_i\right) \qquad var(p_0) = \frac{\sum_{i=1}^{n}(d'_i)^2}{n} - \mu'^2 \tag{C.11}$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\left(\sum_{i=1}^{n} d'_i\right) + \frac{n-2k}{n}x = \mu' + \frac{n-2k}{n}x \tag{C.12}$$

$$\begin{aligned} var(p) = \frac{\sum_{i=1}^{n} d_i^2}{n} - \mu^2 &= \frac{1}{n}\left(\sum_{i=1}^{k}(d'_i - x)^2 + \sum_{i=k+1}^{n}(d'_i + x)^2\right) - \left(\mu' + \frac{n-2k}{n}x\right)^2 \\ &= var(p_0) + \left(1 - (\frac{n-2k}{n})^2\right)x^2 + \frac{2}{n}x\left((\sum_{i=k+1}^{n} d'_i) - (\sum_{i=1}^{k} d'_i) - (n-2k)\mu'\right) \end{aligned} \tag{C.13}$$

$$\frac{var(p) - var(p_0)}{x} = \left(1 - (\frac{n-2k}{n})^2\right)x + \frac{2}{n}\left((\sum_{i=k+1}^{n} d'_i) - (\sum_{i=1}^{k} d'_i) - (n-2k)\mu'\right) \tag{C.14}$$

Let $x \to 0$, we have

$$lim_{x\to 0}\frac{var(p)-var(p_0)}{x} = \frac{2}{n}\left((\sum_{i=k+1}^{n} d_i') - (\sum_{i=1}^{k} d_i') - (n-2k)\mu' \right) \tag{C.15}$$

Proof of Proposition 3. We consider both directions.

a. Suppose p_0 is a local MV of T then by Eq. C.15

$$\begin{aligned} (\sum_{i=k+1}^{n} d_i') - (\sum_{i=1}^{k} d_i') - (n-2k)\mu' = 0 \\ \implies n\sum_{i=k+1}^{n} d_i' - n\sum_{i=1}^{k} d_i' - (n-2k)\sum_{i=1}^{n} d_i' = 0 \\ \implies \frac{1}{k}\sum_{i=1}^{k} d_i' = \frac{1}{n-k}\sum_{i=k+1}^{n} d_i' \end{aligned} \tag{C.16}$$

Thus, p_0 is also a balance point, which completes one direction of Proposition 3.

b. Suppose p_0 is a balance point of T; then,

$$\frac{1}{k}\sum_{i=1}^{k} d_i' = \frac{1}{n-k}\sum_{i=k+1}^{n} d_i' = \mu' \tag{C.17}$$

Substituting $\sum_{i=k+1}^{n} d_i'$ and $\sum_{i=1}^{k} d_i'$ in Eq. C.15 gives

$$lim_{x\to 0}\frac{var(p)-var(p_0)}{x} = ((n-k)-k-(n-2k))\mu' = 0 \tag{C.18}$$

which means, p_0 is a local MV. This completes the proof for Proposition 3. □

Proof of Lemma 6. Suppose p_0 is not a local MV. By Lemma 5, there is a point p_1 on the same edge or an adjacent edge to p_0 such that $\overrightarrow{p_0p_1}$ is a dominant direction of p_0. Letting $y = d(p_0, p_1)$, replacing p with p_1 in Eq. C.15, we get:

$$\begin{aligned}
lim_{y\rightarrow 0}\frac{var(p_1)-var(p_0)}{y} &= \\
\frac{2}{n^2}\big(n\sum_{i\notin Cld_{p_0}(p_1)} d_i(p_0) - n\sum_{i\in Cld_{p_0}(p_1)} d_i(p_0) - (n-2|p_0\triangleright p_1|)\sum_{i\in L} d_i(p_0)\big) &= \\
\frac{4}{n^2}\big(|p_0\triangleright p_1|\sum_{i\notin Cld_{p_0}(p_1)} d_i(p_0) - (n-|p_0\triangleright p_1|)\sum_{i\in Cld_{p_0}(p_1)} d_i(p_0)\big) &< 0
\end{aligned}$$

where the inequality follows from the fact that $\overrightarrow{p_0p_1}$ is a dominant direction (see Eq. C.3). Because the derivative at p_0 approaching from p_1 is negative, there exist a point p' in a small local neighborhood of p_0 towards p_1 such that $var(p') < var(p_0)$. □

Proofs of Proposition 4 – 7 and Lemma 7

Proof of Lemma 7. For the the edge (u,v) (where $u = p(v)$), let $m_1^u = \frac{1}{|u\triangleright v|}\sum_{i\in Cld_u(v)} d_i(u)$ and $m_2^u = \frac{1}{n-|u\triangleright v|}\sum_{i\notin Cld_u(v)} d_i(u)$, and similarly, $m_1^v = \frac{1}{|v\triangleright u|}\sum_{i\notin Cld_v(u)} d_i(v)$ and
$m_2^v = \frac{1}{n-|v\triangleright u|}\sum_{i\in Cld_v(u)} d_i(v)$.

By definition of dominant direction (Eq. C.3), $m_1^u > m_2^u$ and $m_1^v > m_2^v$. On the other hand, since $m_1^u = m_2^v + e_v$ and $m_2^u = m_1^v - e_v$, we have $0 < m_1^u - m_2^u = m_2^v - m_1^v + 2e_v < 2e_v$. Let p be a point on edge e such that $d(p,u) = x = \frac{m_1^u - m_2^u}{2}$. We have:

$$\frac{1}{|p\triangleright u|}\sum_{i\in Cld_p(u)} d_i(p) = m_1^u - x \quad \text{and} \quad \frac{1}{n-|p\triangleright u|}\sum_{i\notin Cld_p(u)} d_i(p) = m_2^u + x$$

$\frac{1}{|p\triangleright u|}\sum_{i\in Cld_p(u)} d_i(p) - \frac{1}{n-|p\triangleright u|}\sum_{i\notin Cld_p(u)} d_i(p) = m_1^u - m_2^u - 2x = 0$. Thus, p is a balance point of T. □

Proof of Proposition 4. Consider a tree T rooted at r_T. If r_T is a local MV, then the proof is complete. If r_T is not a local MV, by Lemma 5 and Lemma 7, there exists an edge $e_0 = (r_T, v_0)$ such that $\overrightarrow{r_Tv_0}$ is a dominant direction of r_T. If v_0 is a balance point of T, or $\overrightarrow{v_0r_T}$ is a dominant direction of v_0, then by Lemma 7 and Proposition 3, there is a local MV p on e_0.

Otherwise, by Lemma 5, v_0 has a dominant direction $\overrightarrow{v_0v_1}$ associated with edge $e_1 = (v_0, v_1)$. Similar to the previous case, if v_1 is a balance point or $\overrightarrow{v_1v_0}$ is a dominant direction of v_1, then there is a balance point p on e_1. Otherwise, v_1 has a dominant direction $\overrightarrow{v_1v_2}$ associated with edge $e_2 = (v_1, v_2)$.

The process can be continued until we reach an edge $e_k = (v_{k-1}, v_k)$ such that either there is a local MV $p \in e_k$ or v_k is a leaf of T. If v_k is a leaf, then it is obvious that $\overrightarrow{v_kv_{k-1}}$ is a dominant direction of v_k. Recall that $\overrightarrow{v_{k-1}v_k}$ is a dominant direction of v_{k-1}. By Lemma 7 and Proposition 3, there is a local MV point p on e_k.

Thus, we can always find at least one local MV in a tree T (if tree T is finite). This completes the proof of Proposition 4. □

Proof of Proposition 5. (Proof by contradiction) Suppose there exists a tree T with a global MV p_0 that is not a local MV. Let edge $e = (u, v)$ be the edge that contains p_0. Since p_0 is not a local MV, by Lemma 6, there exists a point p such that $var(p) < var(p_0)$, which contradicts the definition of global MV. □

Proof of Proposition 6. On tree T, let p be the global MV and $x = d(p, r)$, w denote the child of r that is on the same side as p, and d_i be the shorthand for $d_i(r)$ (i.e. the distance from r to leaf i *of tree T*). We prove that $x \leq (1-\varepsilon)e_w$, and therefore, $p \in e(r_0, w)$. Note that T_0 and T have the same topology but are different in branch lengths. In this proof we use e_v to denote the length of the edge $(p(v), v)$ *of* T_0.

Follow the lemma condition

$$\varepsilon \leq \frac{e_w}{\frac{n}{n-|w|}h + e_w} \implies \frac{n}{n-|w|}\varepsilon h \leq (1-\varepsilon)e_w \tag{C.19}$$

By Proposition 3 and 5, p is a balance point. Therefore,

$$\frac{1}{|p|}\sum_{i\in Cld(p)}(d_i - x) = \frac{1}{|p|}\sum_{i\in Cld(p)} d_i(p) = \frac{1}{n-|p|}\sum_{i\notin Cld(p)} d_i(p) \tag{C.20}$$

Also,

$$\frac{1}{n-|p|}\sum_{i\notin Cld(p)} d_i(p) \geq \frac{1}{n-|p|}\big(\sum_{i\notin Cld(p)}(d_i) + (n-|w|)x - (|w|-|p|)x\big) \tag{C.21}$$

From Eq. C.20 and C.21, we have

$$\begin{aligned}
&\frac{\sum_{i\in Cld(p)}(d_i - x)}{|p|} \geq \frac{\sum_{i\notin Cld(p)} d_i + (n-|w|)x - (|w|-|p|)x}{n-|p|} \\
\Longrightarrow\ &\frac{\sum_{i\in Cld(p)} d_i}{|p|} - x \geq \frac{\sum_{i\notin Cld(p)} d_i}{n-|p|} + \frac{(n-|w|)-|w|+|p|}{n-|p|}x \\
\Longrightarrow\ &\left(1+\frac{n-|w|-|w|+|p|}{n-|p|}\right)x = \frac{2(n-|w|)}{n-|p|}x \leq \frac{\sum_{i\in Cld(p)} d_i}{|p|} - \frac{\sum_{i\notin Cld(p)} d_i}{n-|p|}
\end{aligned}$$

Recall that under our model, T_0 is an ultrametric tree, so that for each leaf i, $\sum_{v\in path(i,r)} e_v = h$. Also, T was obtained by multiplying each edge of T_0 by a random variable with support $[1-\varepsilon, 1+\varepsilon]$. Thus, $(1-\varepsilon)h \leq d_i = \sum_{v\in path(i,r)} e_v\alpha_v \leq (1+\varepsilon)h$. Therefore,

$$\frac{2(n-|w|)}{n}x \leq \frac{2(n-|w|)}{n-|p|}x \leq 2\varepsilon h \implies x \leq \frac{n}{n-|w|}\varepsilon h \leq (1-\varepsilon)e_w$$

Hence, there exists a child w of r such that the global MV belongs to edge (r,w). □

Proof of Proposition 7. Let D_i be the random variable corresponding to the distribution of $d_i(r)$ and P be a random variable giving the position of the global MV root. Then,

$$\begin{aligned}
E[D_i] &= E[\sum_{v\in path(i,r)} e_v\alpha_v] = \sum_{v\in path(i,r)} E[e_v\alpha_v] \\
&= \sum_{v\in path(i,r)} e_v E[\alpha_v] = \sum_{v\in path(i,r)} e_v = h
\end{aligned} \tag{C.22}$$

By the global balance property of P, we can compute

$$X = \frac{1}{2}\left(\frac{\sum_{i \in Cld(P)} D_i}{|P|} - \frac{\sum_{i \notin Cld(P)} D_i}{n - |P|}\right) \tag{C.23}$$

and thus,

$$E[X] = \frac{1}{2}\left(\frac{\sum_{i \in Cld(P)} E[D_i]}{|P|} - \frac{\sum_{i \notin Cld(P)} E[D_i]}{n - |P|}\right) = \frac{1}{2}(h - h) = 0 \tag{C.24}$$

□

C.2 Supplementary figures and tables

Arg.	Description	Value for D1	Value for D2
RS	Number of replicates	100	20
RL	Number of loci	500	50
RG	Number of genes	1	
SB	Speciation rate	Log normal(1.0*e*-7,1.0*e*-6)	
SD	Extinction rate	Log normal(1.0*e*-7,SB)	
ST	Maximum tree length	Lognormal(14.41412,1)	Lognormal(16,1)
SL	Number of taxa	30	
SO	Root to crown ratio	R/C	
SI	Number of individuals per species	1	
SP	Global population size	Uniform(10000,1000000)	
SU	Global substitution rate	Log normal(-17.27461,0.6931472)	
HH	Gene by lineage specific locus tree parameter	1	
HS	Species specific branch rate heterogeneity	Log normal(α,1)	
HL	Gene family specific rate heterogeneity	Log normal(1.551533,0.6931472)	
HG	Gene by lineage specific rate heterogeneity	Log normal(α,1)	
CS	Random number generator seed	9644	

Table C.1: Parameters used in SimPhy simulation

Root to crown ratios and Divergence from the strict clock are shown with variables α and R/C. These parameters change for each model condition and are available in Table C.2.

Model Condition.	R/C for D1 and D2	α for D1 and D2
1	0	1.5
2	0.25	1.5
3	0.5	1.5
4	1	1.5
5	2	1.5
6	4	1.5
7	1	0.15
8	1	5
9	0	0.15
10	0	5

Table C.2: R/C and α for different model conditions in datasets D1 and D2.

Methods compared	**p-value**		**Mean MS ST error**	
	method	**clock par.**	**1st method**	**2nd method**
STAR True vs STAR Ideal	$< 10^{-5}$	0.0638	7.6313	7.6313
STAR Ideal vs STAR OG	0.5820	0.0041	11.8875	12.0844
STAR Ideal vs STAR MV	0.1892	0.0008	11.8875	13.0938
STAR OG vs STAR MV	0.4768	0.0008	12.0844	13.0938
STAR OG vs NJst	0.1619	0.0085	12.0844	13.5906

ANOVA tests were performed on the D1 (30-taxon) dataset for pairs of methods. Matching-split (MS) error is used as the metric. The tests were performed on the subset of D1 where outgroup exists. For true gene trees, the true root is known. For estimated gene trees, the Ideal is the rooting position that minimizes triplet error to the true gene trees. p-values are shown for the significance of differences between the error of the two methods specified in each row, and for the differences in error among the three levels of clock divergence parameter, respectively.

Table C.3: Species tree estimation accuracy using rooted and unrooted gene trees

C.3 Supplementary methods

C.3.1 Simulation setup

In order to simulate the gene sequences we used Indelible for datasets D1 and D2, with sequence lengths and mutation parameters drawn randomly from distributions described below. D1 has 30 taxa and D2 is a large dataset with 2000 or 5000 taxa. Note that in order to match the level of gene tree error observed in D1 in the D2 dataset, which included many more species, we

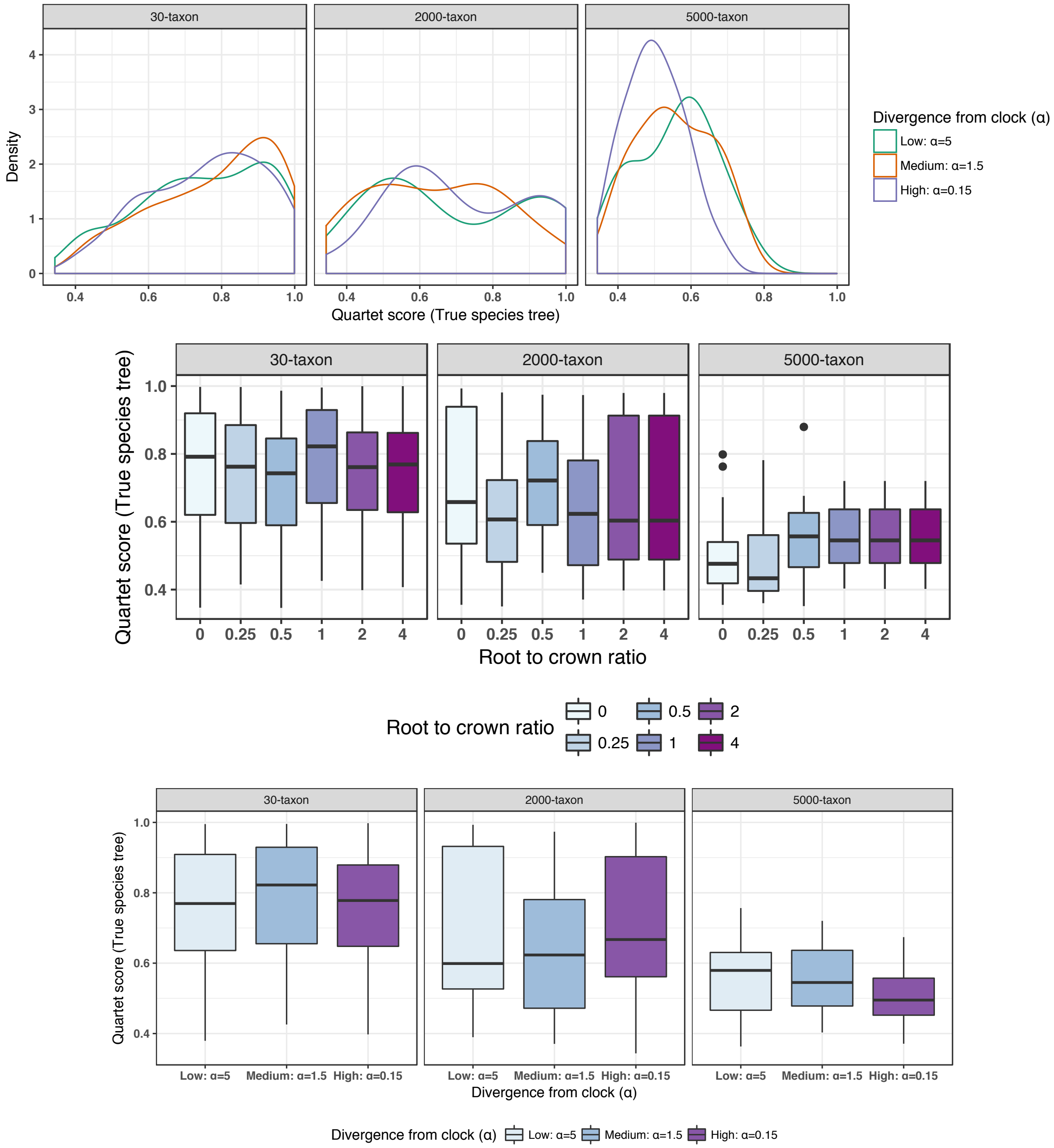


Figure C.2: ILS levels for new simulated datasets D1 and D2. Density plots (top) and box plots (middle and bottom) are shown for the quartet score of the true species tree with respect to the true gene trees, as a measure of the amount of ILS. Top: R/C=1. Middle: divergence from the clock = 1.5. Bottom: R/C=1.

set our sequence length hyperparameters such that we had longer sequence lengths in D2.

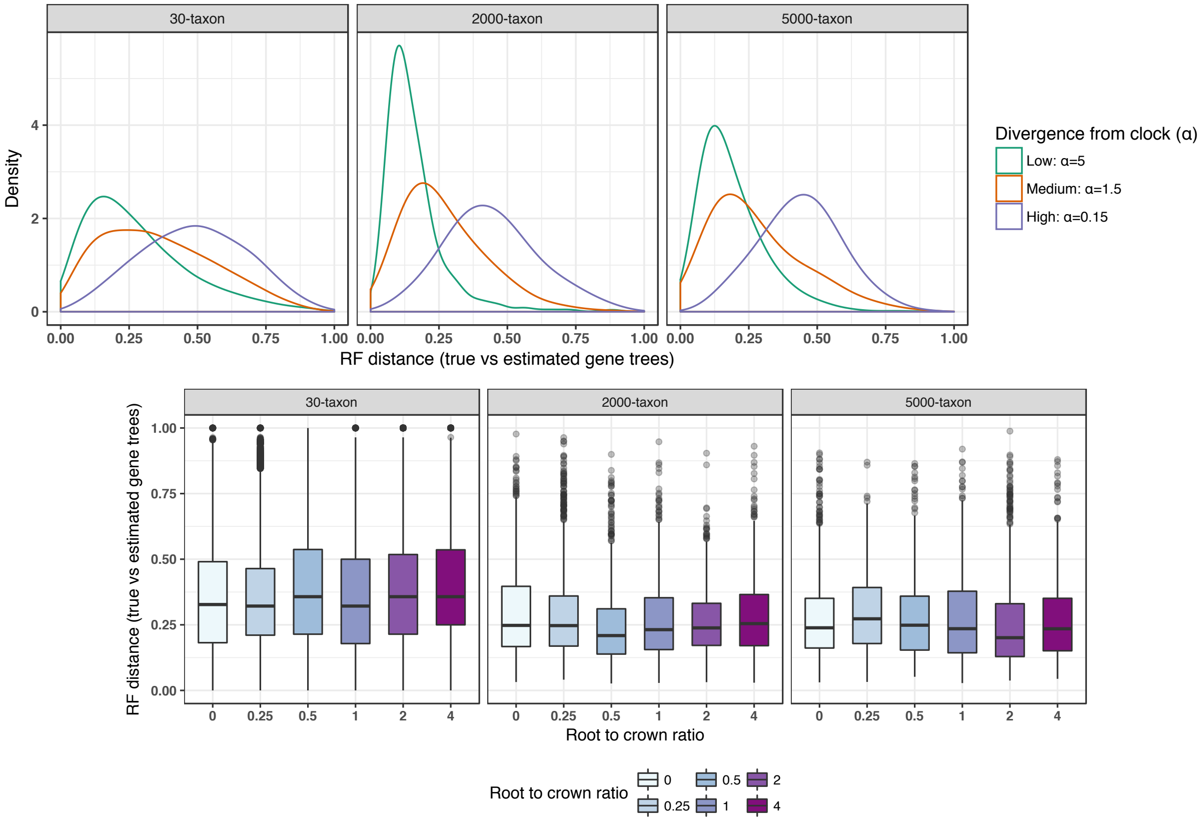


Figure C.3: Gene tree estimation error for datasets D1 and D2. The normalized RF distance is shown between true gene trees and the estimated gene trees. Top: density plots with R/C= 1; Bottom: boxplots with the divergence clock parameter set to 1.5.

Gene lengths: In D1, for each gene, we sample the sequence lengths from a log normal distribution. The parameters of the log normal (μ and σ) are drawn randomly from gamma and uniform distributions, respectively, for each individual replicate. We draw μ from a distribution because we want some replicates with high gene length (thus, low gene tree error) and some with low gene length. Similarly, we draw σ from a distribution to have replicates with high or low gene tree error variation.

Our goal was to have an average gene length of roughly 450 sites long across all datasets, which would lead to reasonable average levels of gene tree error. The σ parameter was drawn from a uniform random variable between (0.3,0.7) with the average of 0.5, and this range was empirically derived by trial and error. The mean of log-normal distribution is given by $e^{\mu+\sigma^2/2}$.

For this number to be around 450, we need $\mu + \sigma^2/2 = \log(450)$. Replacing σ with its expected value, 0.5, we get that the expected value of μ should be $log(450) - 1/8$. The gamma distribution (which we use for μ) has an expected value of $shape \times scale$. We empirically observed that a scale of 0.033 results in sufficient variations. So in order to have the mean 450 for log-normal, we parameterize the gamma distribution with scale 0.033 and the shape $(log(450) - 1/8)/0.033$ and draw a value X from this distribution. This procedure gives us a left-skewed distribution with many numbers below 450. In order to make the distribution right-skewed (and avoid many genes with very few sites), we used a simple trick. We use $Y = 2\log(450) - 1/8 - X$ as our draw of μ. The expected value of Y remain $\log(450) - 1/8$, which in turn, leads to expected gene length of 450; however, the distribution becomes right-skewed. This gives us an empirical average sequence length of 495. The median sequence lengths is between 370 and 422 in 90% of replicates.

In D2, for each gene, we used the same strategy but with a target gene length of 700bp instead of 450bp (since larger trees need more sites to achieve similar accuracy). The rest of the procedure remains the same. The empirical average sequence length was 766, and the median sequence lengths was between 294 and 1236 in 90% of replicates.

Base frequencies: For both datasets D1 and D2 we used a Dirichlet(36 26 28 32) to draw base frequencies for A, C, G, and T. These values are ML estimates of the three previously published large biological datasets, and are obtained from a previous dataset [219].

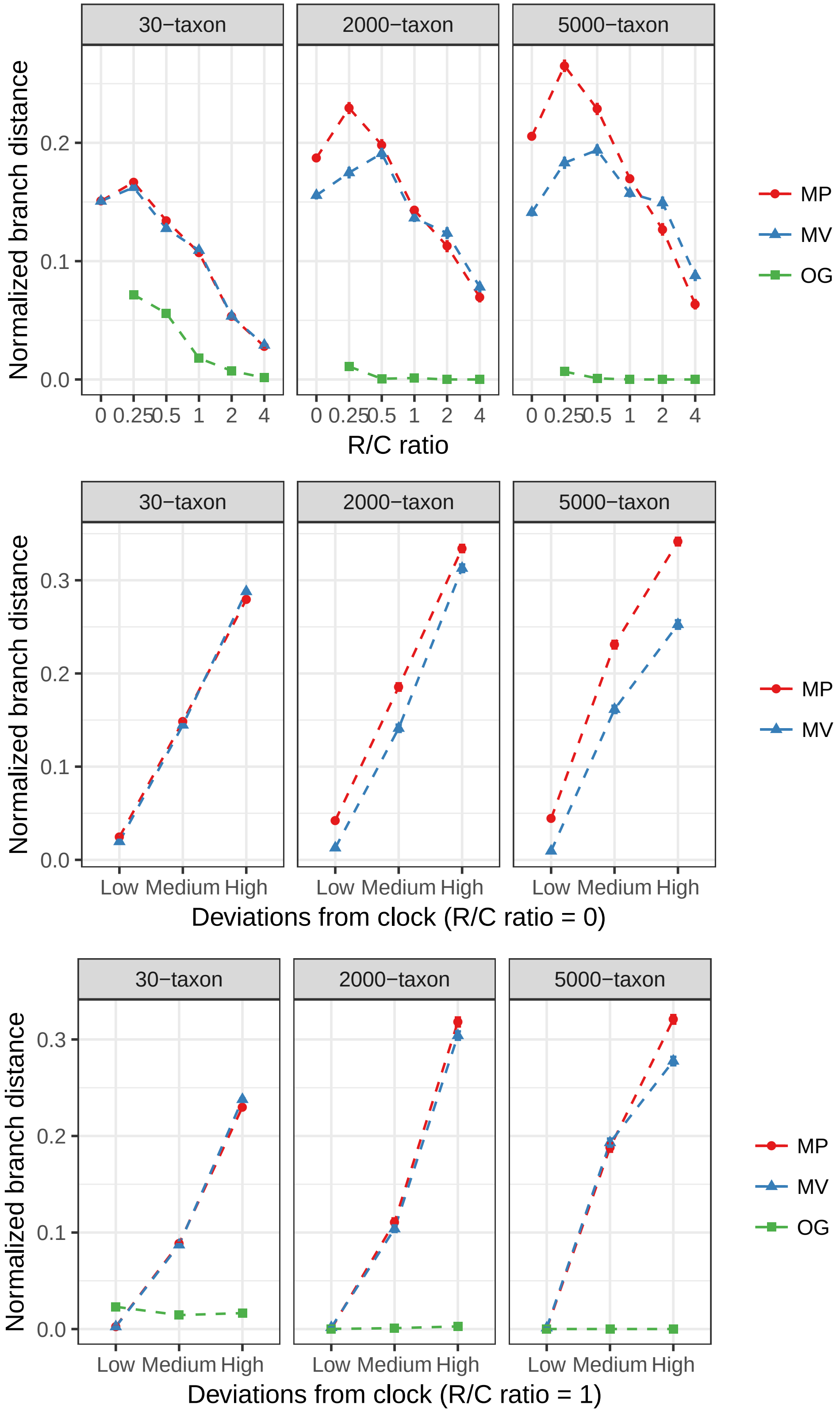


Figure C.4: Normalized branch distance in true rooted gene trees for datasets D1 and D2. The number of branches away from the true root is normalized by the tree depth and is shown for all three methods of rooting.

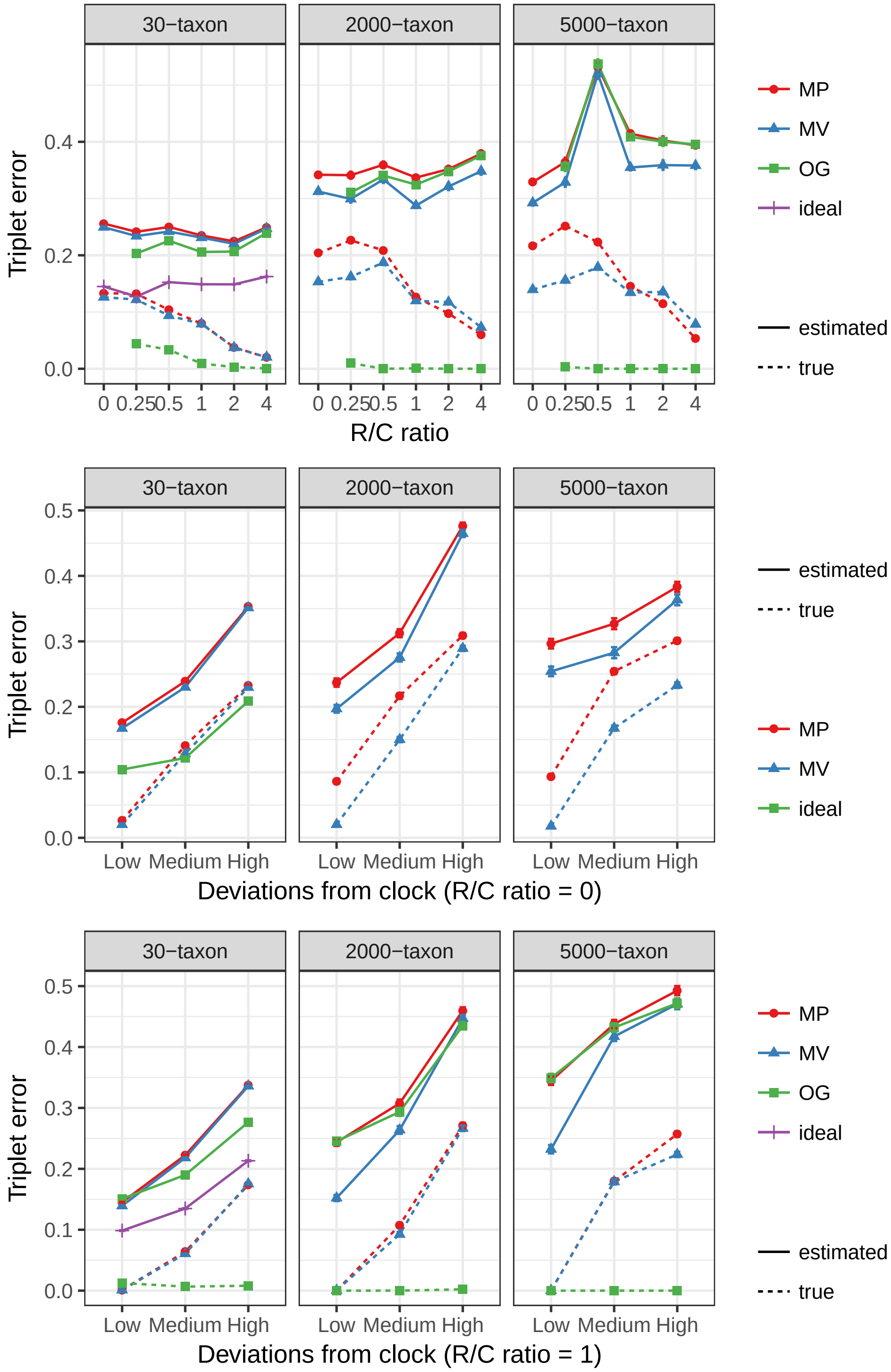


Figure C.5: Triplet error in true and estimated rooted gene trees for datasets D1 and D2. Absolute triplet distance is shown for all three methods of rooting plus the ideal rooting for D1 where a brute force calculation was feasible (the rooting that minimizes the triplet distance to the true tree).

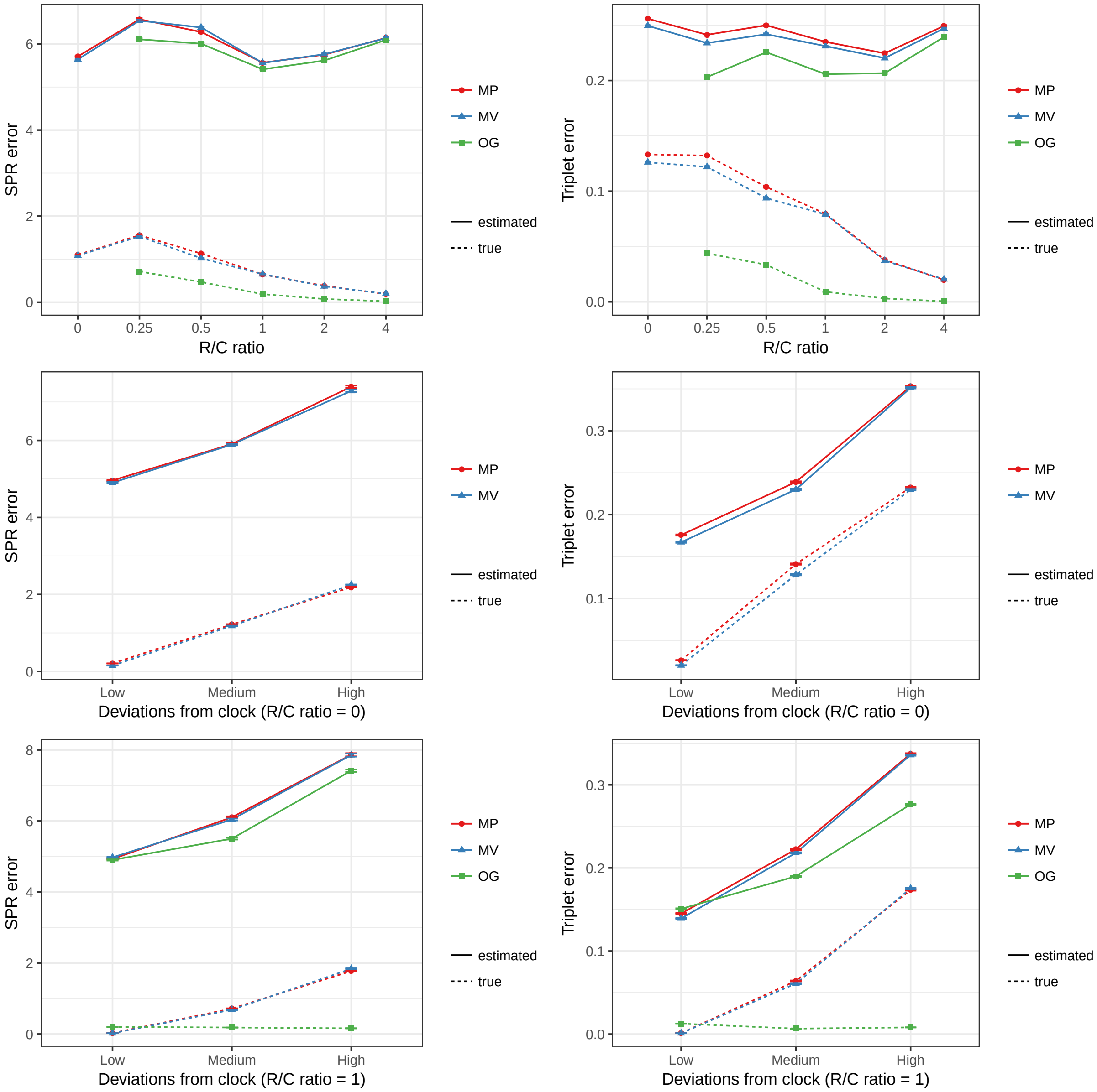


Figure C.6: SPR and Triplet error in true and estimated rooted gene trees for the 30-taxon dataset where SPR computation is feasible. Top: SPR and Triplet error with different R/C ratio. Middle and Bottom: SPR and Triplet error with different levels of deviations

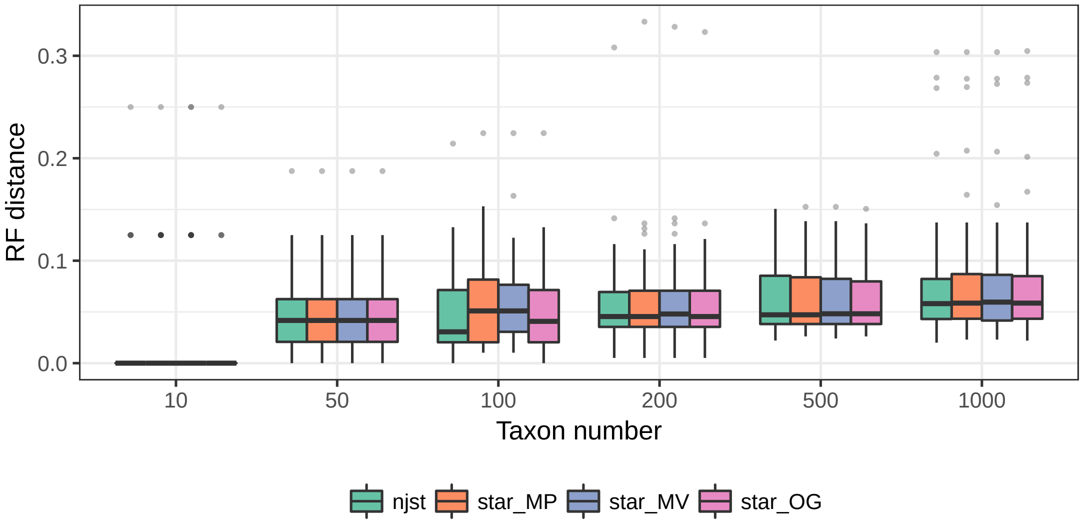


Figure C.7: STAR and NJst error on estimated gene trees for dataset D3. Species trees are estimated on estimated gene trees. RF distance is shown for NJst and STAR with all three methods of rooting.

Appendix D

Supplementary materials for Chapter 4

D.1 Supplementary text

D.1.1 Skewness of the penalty terms of LSD and LogDate under Gamma clock model

Suppose the mutation rates μ_i are drawn i.i.d. from a Gamma distribution $\Gamma(\alpha, \beta)$ where α and β are the shape and rate parameters and there is no branch estimation error (i.e. $\hat{b}_i = b_i = \mu_i \tau_i$ for all branch i). Then the mean of μ_i is $\mu = \frac{\alpha}{\beta}$. Define the rate multipliers as $r_i = \frac{\mu_i}{\mu}$, then we have $r_i \sim \Gamma(\alpha, \alpha)$. Recall that the penalty terms of LSD is $\frac{\mu\tau_i}{\hat{b}_i} - 1 = \frac{\mu\tau_i}{\mu_i\tau_i} - 1 = \frac{1}{r_i} - 1$ and the penalty terms of LogDate is $\log r_i$. Note that $\frac{1}{r_i}$ follows an inverse Gamma distribution (with shape α and scale α) and $\log r_i$ a Log-Gamma distribution. Therefore, the skewness of the penalty terms can be computed for LSD to be $\frac{4\sqrt{\alpha-2}}{\alpha-3}$ and for LogDate to be $\frac{\psi^{(3)}(\alpha)}{[\psi^{(2)}(\alpha)]^{3/2}}$ (where $\psi^{(2)}$ and $\psi^{(3)}$ are the digamma and trigamma functions, respectively. Figure D.1 shows the skewness of LSD and LogDate when the rate's variance increases.

Algorithm 5 Setup the linear constraints Ψ for tree T given a set of calibration points.

```
function SETUPCONSTRAINTS(T)
    $\Psi \leftarrow \{\}$
    for $w$ in post-order traversal of $T$ do
        if $w$ is a leaf then
            if $w$ is calibrated then
                nearest_timepoint($w$) $\leftarrow w$
            else
                nearest_timepoint($w$) $\leftarrow \emptyset$
        else
            $(w_1, w_2) \leftarrow$ Children($w$)
            $u \leftarrow$ nearest_timepoint($w_1$)
            $v \leftarrow$ nearest_timepoint($w_2$)
            if $w$ is calibrated then
                nearest_timepoint($w$) $\leftarrow$ w
                if $u! = \emptyset$ then
                    $\Psi \leftarrow \Psi \cup \psi(w, u)$
                if $v! = \emptyset$ then
                    $\Psi \leftarrow \Psi \cup \psi(w, v)$
            else
                if $u == \emptyset$ then
                    nearest_timepoint($w$) $\leftarrow v$
                else
                    if $v == \emptyset$ then
                        nearest_timepoint($w$) $\leftarrow u$
                    else
                        nearest_timepoint($w$) $\leftarrow u$ if $(d(w, u) < d(w, v))$ else $v$
                        $\Psi \leftarrow \Psi \cup \psi(u, v)$
```

Claim 6. The optimal $\mathbf{x}^*$ in Eq. 4.5 yields a set $\nu = \{\nu_1, \nu_2, \ldots, \nu_{2n-1}\}$ that has the maximum joint probability under a model of rate variation where ν_i are i.i.d, $\nu_i \sim LogNormal(\mu_0, \sigma_0^2)$, and $Mode(\nu_i) = 1$ for any μ_0 and σ_0^2, subject to the constraints Ψ.

Proof. We have: $\nu_i \sim LogNormal(\mu_0, \sigma_0^2) \implies Mode(\nu_i) = e^{\mu_0 - \sigma_0^2} \implies 1 = e^{\mu_0 - \sigma_0^2} \implies \mu_0 = \sigma_0^2$. In other words, for the conditions with mode 1, we only have one free parameter.

The logarithm of the joint probability of ν under the LogNormal model of rate variation can be written as follows:

$$\begin{aligned} P(\nu_1, \nu_2, \ldots, \nu_{2n-1} | \mu_0, \sigma_0^2) &= \textstyle\sum_{i=1}^{2n-2} \log\left(\frac{1}{\nu_i \sigma_0 \sqrt{2\pi}} \exp\left(-\frac{(\log \nu_i - \mu_0)^2}{2\sigma_0^2}\right)\right) \\ &\propto \textstyle\sum_{i=1}^{2n-2} \left(-\log \nu_i - \frac{\log^2 \nu_i - 2\mu_0 \log \nu_i}{2\mu_0}\right) \\ &= \textstyle\sum_{i=1}^{2n-2} -\log^2 \nu_i \end{aligned} \tag{D.1}$$

Thus, maximizing the joint probability of ν is equivalent to minimizing Eq. 4.5, subject to the constraints Ψ. □

Lemma 8. *The length of the shortest path from the root of a binary tree to its leaves is at most* $\log n$ *where n is the number of leaves in the tree.*

Proof. Consider a rooted binary tree $\mathcal{T}$ with n leaves; let r be the root and h be the length of the shortest path from r to the leaves of $\mathcal{T}$. We need to prove that $h \leq \log_2 n$.

Let $\mathcal{D}_i$ be the set of nodes in $\mathcal{T}$ with depth i, that is, $\mathcal{D}_i = \{w \in \mathcal{T} | d(r, w) = i\}$. We first prove that $|\mathcal{D}_i| = 2^i \; \forall i \leq h$ where $|\mathcal{D}_i|$ denotes the cardinality of $\mathcal{D}_i$. We prove this by induction. The base case $i = 0$ holds since the root r is the only node with depth 0. Suppose we have $|\mathcal{D}_k| = 2^k$ and $k < h$, we need to prove that if $k + 1 \leq h$ then $|\mathcal{D}_{k+1}| = 2^{k+1}$. Note that a node $v \in \mathcal{D}_{k+1}$ if and only if its parent $par(v) \in \mathcal{D}_k$. Because $\mathcal{T}$ is a binary tree, each node in $\mathcal{T}$ must either has no child (leaf node) or two children (internal node). Since $k < h$, there must be no leaf node in $\mathcal{D}_k$, otherwise, a leaf v in $\mathcal{D}_k$ has $d(r, v) = k < h$, which defines a root-to-leaf path that is shorter than h and contradicts the definition of h. Thus, each node in $\mathcal{D}_k$ has exactly 2 children, making $|\mathcal{D}_{k+1}| = 2 * |\mathcal{D}_k| = 2 * 2^k = 2^{k+1}$.

Now we have $|\mathcal{D}_h| = 2^h$. To prove that $h \leq \log_2 n$, note that $\mathcal{D}_h$ contains a mixture of leaves and internal nodes and each internal node in $\mathcal{D}_h$ must have more than one leaf below it. Therefore, the size of $\mathcal{D}_h$ is at most the size of the leaf set of $\mathcal{T}$; that is, $|\mathcal{D}_h| \leq n$. Thus, we have $2^h = |\mathcal{D}_h| \leq n \implies h \leq \log_2 n$. □

Claim 7. If all the leaves have sampling times and there is no other calibration points given for internal nodes, the matrix corresponding to the constraints Ψ setup by Algorithm 5 has $O(n \log(n))$ non-zero elements, where n is the number of leaves in the input tree $\mathcal{T}$.

Proof. Let $\mathcal{P}(w)$ denote the shortest path from a node w to its leaves and let $|\mathcal{P}(w)|$ denote the length of this path. Let $\mathcal{T}_w$ be the clade of $\mathcal{T}$ below w and let $|w|$ denote the size of this clade (i.e. the number of leaves below w). Applying lemma 8 on $\mathcal{T}_w$, we have $|\mathcal{P}(w)| \leq \log_2 |w| \leq \log_2 n$ for all $w \in \mathcal{T}$.

Note that if all leaves have sampling times, Algorithm 5 adds exactly one constraint for each internal node in the tree. For each node w with two children $c_l(w)$ and $c_r(w)$, the non-zero elements of the constraint added when node w is visited must locate on $\mathcal{P}(c_l(w))$, $\mathcal{P}(c_r(w))$, and the two branches $(w, c_l(w))$ and $(w, c_r(w))$. Let η_w denote the number of non-zero elements of the constraint defined by node w, then $\eta_w \leq |\mathcal{P}(c_l(w))| + |\mathcal{P}(c_r(w))| + 1 + 1 \leq 2\log_2 n + 2$. Thus, the total number of non-zeros in all constraints corresponding to the $n-1$ internal nodes is bounded above by $(n-1)(2\log_2 n + 2) \in O(n \log n)$. □

D.2 Hybrid rate Angiosperm

Beaulieu *et al.*([23]) simulated a hybrid rate model for a phylogeny of seed plants in which evolutionary rates formed local clocks in certain clades of the tree. The authors simulated that data in 5 scenarios where they change the relative ratios between some clades in the tree, as follow:

- scenario 1 = 3:1 herbaceous to woody
- scenario 2 = 6:1 herbaceous to woody
- scenario 3 = 4:1 angio. to gymno.; 3:1 herbaceous to woody angio.
- scenario 4 = 4:1 angio. to gymno.; 3:1 herbaceous to woody angio.; Gnetales herbaceous angio.
- scenario 5 = 4:1 angio. to gymno.; 3:1 herbaceous to woody angio.; Gnetales woody angio.

The time tree and 100 simulated phylograms for each of these five scenarios were downloaded from the Dryad Repository provided by the authors. We used the provided phylograms to estimate the time tree using wLogDate, RelTime, and LF and compare the estimated age of Angiosperm to the true tree. Without the simulated sequences, we could not run BEAST. However, we show the BEAST results reported by the original study ([23]). We aware that the comparison to BEAST must be made with cautions, because the experimental settings were different as we will state below:

- As RelTime cannot run without outgroups, we had to use the 20 species on the clade outside the Angiosperm as outgroups. As such, this entire clade is ignored by RelTime and only 91 species out of 111 are included in the time tree. We used the same setting for wLogDate and LF. Because of this fact, 5 calibration points belong to the 20 species in the outgroups are also discarded out of the total 20 calibration points. However, in their original study, the authors ran BEAST using 20 calibration points (instead of 15 points) to date the full tree with 111 species (instead of 91 species).

- In their original study, the authors gave BEAST a distribution instead of exact-time for each calibration, as opposed to the exact-time points as we used to run LF, RelTime, and wLogDate.

Supplementary figures and tables

Tree Model	Clock Model	BEAST_strict	BEAST_lnorm	LF	LSD	RTT	wLogDate
M1	Lognormal	0.0007	0.0011	**0.0004**	0.0005	0.0007	**0.0004**
	Gamma	0.0009	0.0014	0.0005	**0.0004**	0.0009	0.0006
	Exponential	0.0009	0.0022	**0.0008**	0.0013	0.0013	0.0009
M2	Lognormal	**0.0005**	0.0014	**0.0005**	**0.0005**	0.0008	**0.0005**
	Gamma	**0.0006**	0.0013	0.0007	0.0007	0.0009	0.0007
	Exponential	**0.0013**	0.0038	0.0015	0.0020	0.0019	0.0015
M3	Lognormal	**0.0003**	**0.0003**	0.0006	0.0004	0.0008	0.0006
	Gamma	**0.0003**	**0.0003**	0.0006	0.0004	0.0006	0.0006
	Exponential	0.0010	0.0011	**0.0009**	0.0027	0.0012	**0.0009**
M4	Lognormal	**0.0007**	0.0008	0.0008	**0.0007**	0.0010	0.0008
	Gamma	**0.0006**	0.0007	0.0008	0.0008	0.0010	0.0007
	Exponential	0.0020	**0.0016**	**0.0016**	0.0037	0.0018	0.0017
Average		**0.0008**	0.0013	**0.0008**	0.0012	0.0011	**0.0008**

Table D.1: Mean absolute error of the inferred mutation rate of BEAST_strict, BEAST_lognorm, LF, LSD, and wLogDate.

Tree Model	Clock Model	BEAST_strict	BEAST_lnorm	LSD	LF	wLogDate
M1	Lognormal	2784.01	6018.80	15.01	17.47	73.65
	Gamma	2823.09	6082.62	17.81	20.29	83.19
	Exponential	2696.61	5840.18	17.44	19.40	723.61
M2	Lognormal	2425.83	5207.20	18.61	19.94	39.32
	Gamma	2466.24	5303.63	20.12	21.47	81.14
	Exponential	2385.73	5169.07	20.87	22.12	418.01
M3	Lognormal	3848.01	8204.00	35.25	38.07	55.92
	Gamma	3850.71	8211.09	38.71	41.55	65.51
	Exponential	3826.12	6520.19	40.10	43.05	280.55
M4	Lognormal	2914.03	6201.59	28.84	30.40	38.92
	Gamma	2901.59	6184.01	30.78	32.29	41.56
	Exponential	2855.79	6145.03	32.96	34.54	371.97

Table D.2: Average running time (seconds) of BEAST_strict, BEAST_lognorm, PhyML + LF, PhyML + LSD, and PhyML + wLogDate with 10 initials on simulated data.

Data	LSD	LF	wLogDate
H1N1 (n=892)	< 1 sec	1 min	3 mins
HIV San Diego (n=904)	< 1 sec	11 mins	24 mins
Ebola (n=1610)	2 secs	4 mins	3 mins

Table D.3: Running time of LSD, LF, and wLogDate on biological datasets.

Tree Model	Clock Model	$w_i = 1$	$w_i = \hat{b}_i + \tilde{b}$	$w_i = (\hat{b}_i + \tilde{b})^2$	$w_i = \log(1 + \hat{b}_i + \tilde{b})$	$w_i = \sqrt{\hat{b}_i + \tilde{b}}$	$w_i = \log(1 + \sqrt{\hat{b}_i + \tilde{b}})$
M1	Lognormal	0.020	0.019	0.027	0.019	**0.018**	**0.018**
M1	Gamma	0.020	0.018	0.026	0.018	**0.017**	**0.017**
M1	Exponential	0.363	0.051	0.062	0.048	**0.037**	0.039
M2	Lognormal	0.107	0.043	0.062	0.043	**0.038**	0.042
M2	Gamma	0.100	0.046	0.064	0.046	**0.043**	0.044
M2	Exponential	0.359	0.135	0.166	0.129	**0.099**	0.100
M3	Lognormal	**0.039**	0.051	0.134	0.050	**0.039**	**0.039**
M3	Gamma	0.041	0.050	0.132	0.049	**0.039**	**0.039**
M3	Exponential	0.220	0.173	0.349	0.168	0.105	**0.103**
M4	Lognormal	0.074	0.098	0.172	0.085	**0.070**	0.071
M4	Gamma	0.098	0.086	0.172	0.082	0.070	**0.069**
M4	Exponential	0.578	0.397	1.042	0.349	0.301	**0.283**

Table D.4: Average RMSE of the internal node ages inferred by different weight functions for LogDate. Numbers are rounded to the closest 3 decimal digits. Recall that $\hat{b}_i$ is the estimated branch length and $\tilde{b}$ is a small constant.

Tree Model	MCMC Chain	Relative Error	Run Time (hours)
M1	10^7	1.00	1.66
M2	5×10^7	0.99	7.6
M3	5×10^7	0.99	11.3
M4	20×10^7	0.98	35.6

Table D.5: BEAST convergence analysis: BEAST was run on Lognormal clock models with the correct prior. For each tree model, we run BEAST with a sufficiently long MCMC chain to ensure the effective-sample-size (ESS) of all parameters are at least 200. We report the length of the MCMC chain, relative error of node age estimates with respect to BEAST using 10 millions MCMC chain, and the running time.

Tree Model	Posterior ESS	Likelihood ESS	MeanRate ESS	RootHeight ESS
M1	420.5	753.3	382.0	406.9
M2	766.9	4224.4	430.1	647.3
M3	1242.3	4024.1	531.0	430.3
M4	1254.9	16336.7	744.1	847.2

Table D.6: BEAST convergence analysis: BEAST was run on Lognormal clock models with the correct prior. For each tree model, we run BEAST with a sufficiently long MCMC chain to ensure the effective-sample-size (ESS) of all parameters are at least 200. We report the average ESS of posterior, likelihood, rootHeight, and meanRate of the first 10 replicates of each tree model.

Replicate	LF	RelTime	wLogDate
1	**0.10**	0.26	0.11
2	0.12	0.21	**0.10**
3	0.11	**0.09**	0.10
4	**0.08**	0.10	0.22
5	**0.05**	0.26	0.06
6	0.10	**0.06**	0.08
7	0.09	0.10	**0.08**
8	0.07	**0.06**	**0.06**
9	0.09	0.08	**0.07**
10	0.17	0.08	**0.07**
Average	0.10	0.13	**0.09**

Table D.7: Comparison of LF, RelTime, and wLogDate on autocorrelated rate dataset. The Root-mean-square error (RMSE) of un-calibrated internal node ages is normalized by the tree height and reported for each replicate. Results discarded the two tests where LF produced extremely erroneous time tree. Refer to Fig. D.9 for a complete picture.

Replicate	DAMBE	wLogDate
1	**7.69**	11.54
2	13.25	**10.88**
3	9.04	**8.30**
4	9.37	**9.10**
5	**3.78**	4.09
6	**4.86**	4.92
7	13.98	**12.46**
8	6.91	**6.43**
9	13.36	**10.93**
10	**14.81**	15.88
Average	9.66	**9.40**

Table D.8: Average relative error (%) of DAMBE and wLogDate in estimating unit time trees on the autocorrelated rate model. For each of the 438 internal nodes across the 10 simulated trees, the relative errors of the inferred divergence times by DAMBE and wLogDate to that of the true normalized time tree are computed. The average error of all nodes per tree replicate and the average error of all 438 nodes are shown for each method.

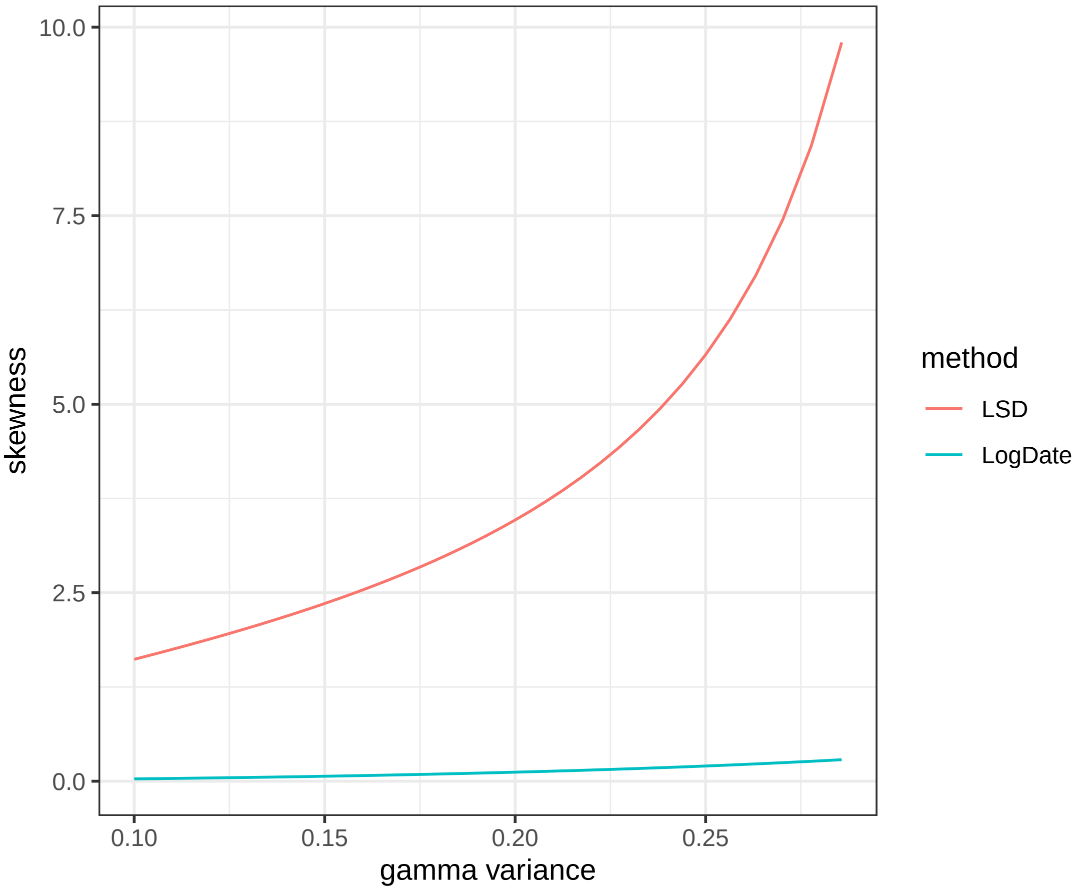


Figure D.1: The skewness of the LSD and LogDate penalty terms when the rate multipliers r_i are drawn i.i.d. from a Gamma distribution with different α. The x-axis shows the variance of r_i and the y-axis shows the skewness of the penalty terms of LSD and LogDate.

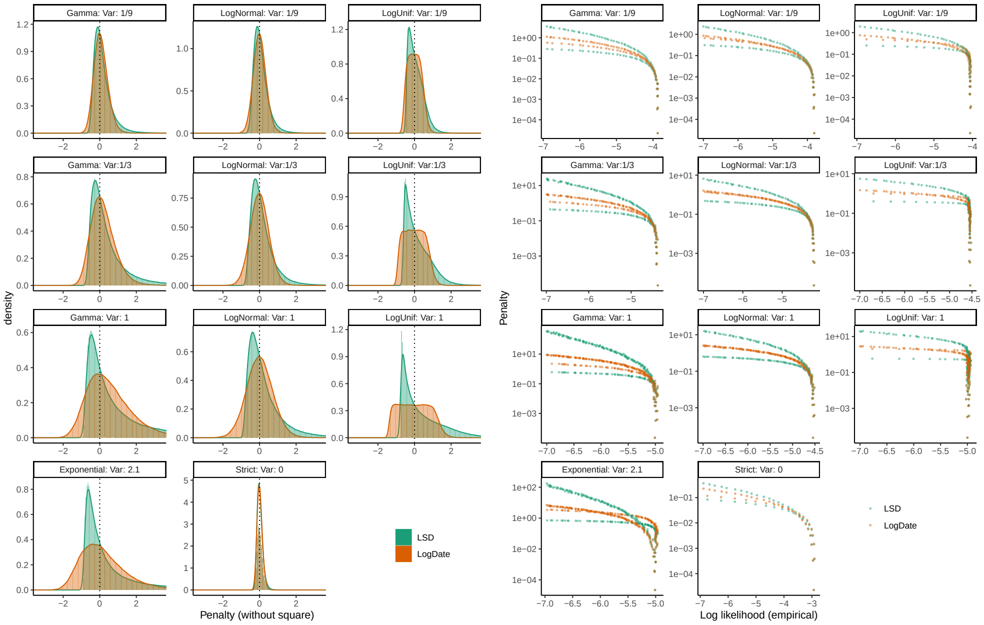
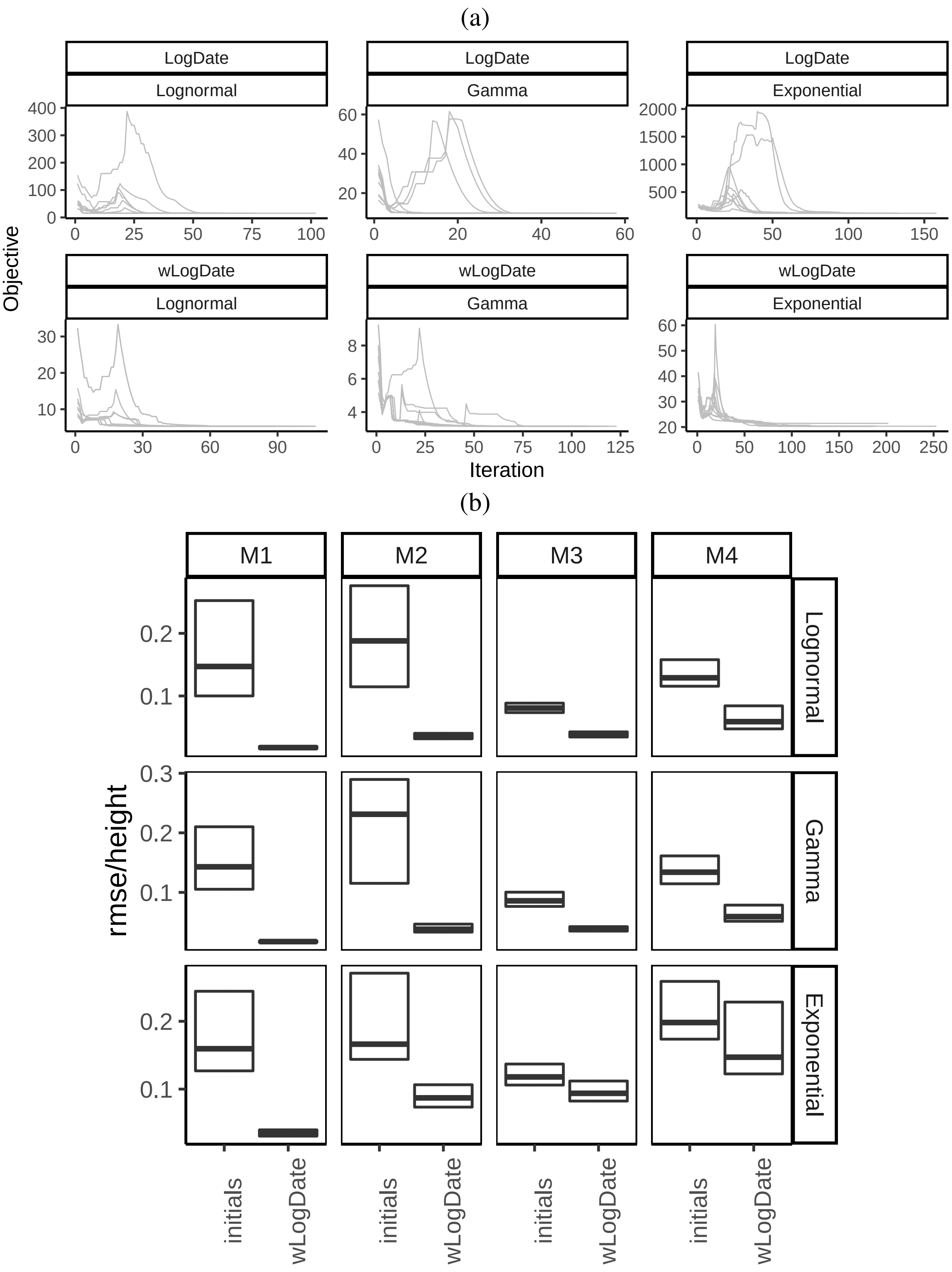


Figure D.2: Left: The density and histograms of the penalty terms (without square) used by LSD ($b_i/\hat{b}_i - 1$) and LogDate ($\log b_i/\hat{b}_i$) under different clock models. Fixing $\mu\tau_i = 0.1$, we draw 500000 values for r_i from a LogNormal, Gamma, or Exponential distribution with median equal to 1 and variance equal to 1/9, 1/3, 1, or 2.1. To simulate strict clock, we fixed $r_i = 1$. We then simulate estimated branch length for each replicate following the [340] model, by drawing $\hat{b}_i$ from a normal distribution with mean $b_i = r_i\mu\tau_i$ and variance b_i/s. Right: The penalty of LSD and LogDate versus the empirical log-likelihood of $\hat{b}_i$ for the models described above. To compute the empirical likelihood, we divide $\hat{b}$ observations into small bins and the empirical likelihood of each bin is estimated as the frequency of the data assigned to it. Ideally, increasing likelihood should monotonically decrease penalty. LogDate is closer to this idea than LSD across all models, especially with higher variance of r_i.


(a)

(b)

Figure D.3: (a) Objective value versus iteration of the LogDate and wLogDate runs on one arbitrarily selected simulated tree (M4, replicate 2). Each of the two methods were run using 10 random initial points generated using the strategy described in the main text. (b) Normalized root-mean-square error of wLogDate versus the 10 initials used to run wLogDate.

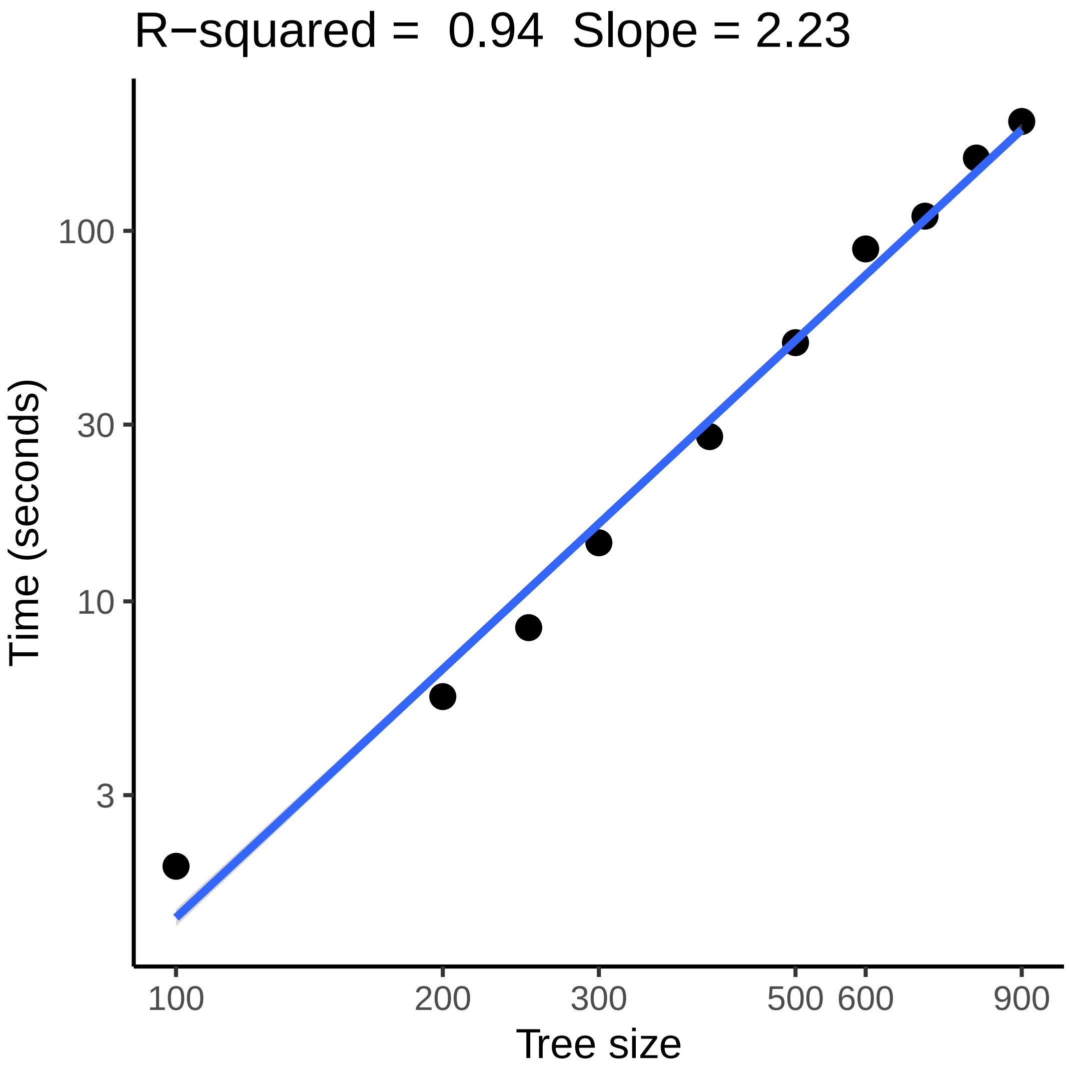


Figure D.4: Running time of wLogDate on random subsets of the HIV dataset. For each tree size, wLogDate was run 100 times on 10 random subsets each with 10 initial points. Each dot represents the average run time of wLogDate per subset per initial point. Both axes are scaled in log (base 10). The slope of the line (2.23) shows the polynomial degree of the running time increase of wLogDate. Thus, wLogDate scales slightly worse than quadratically with increased numbers of species.

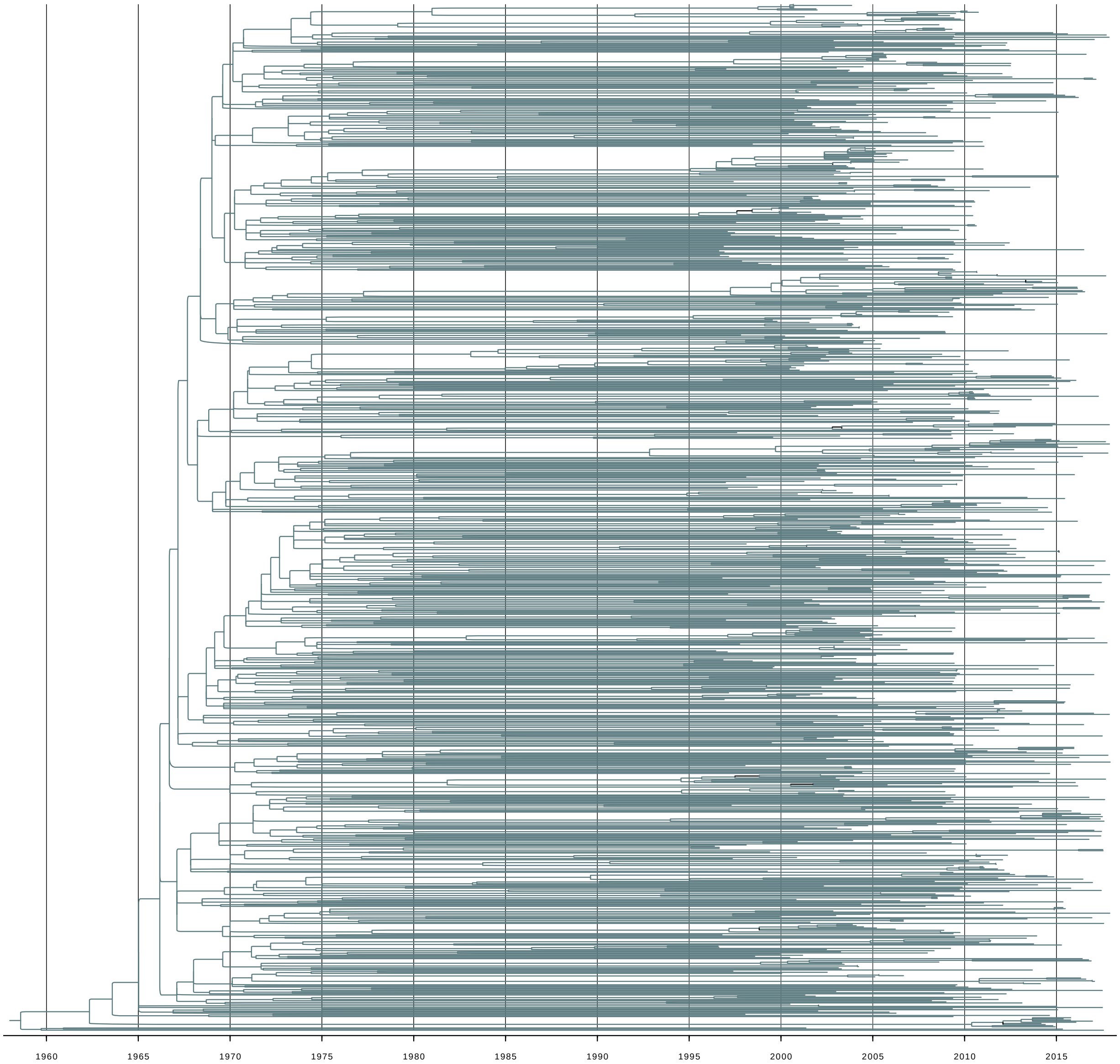
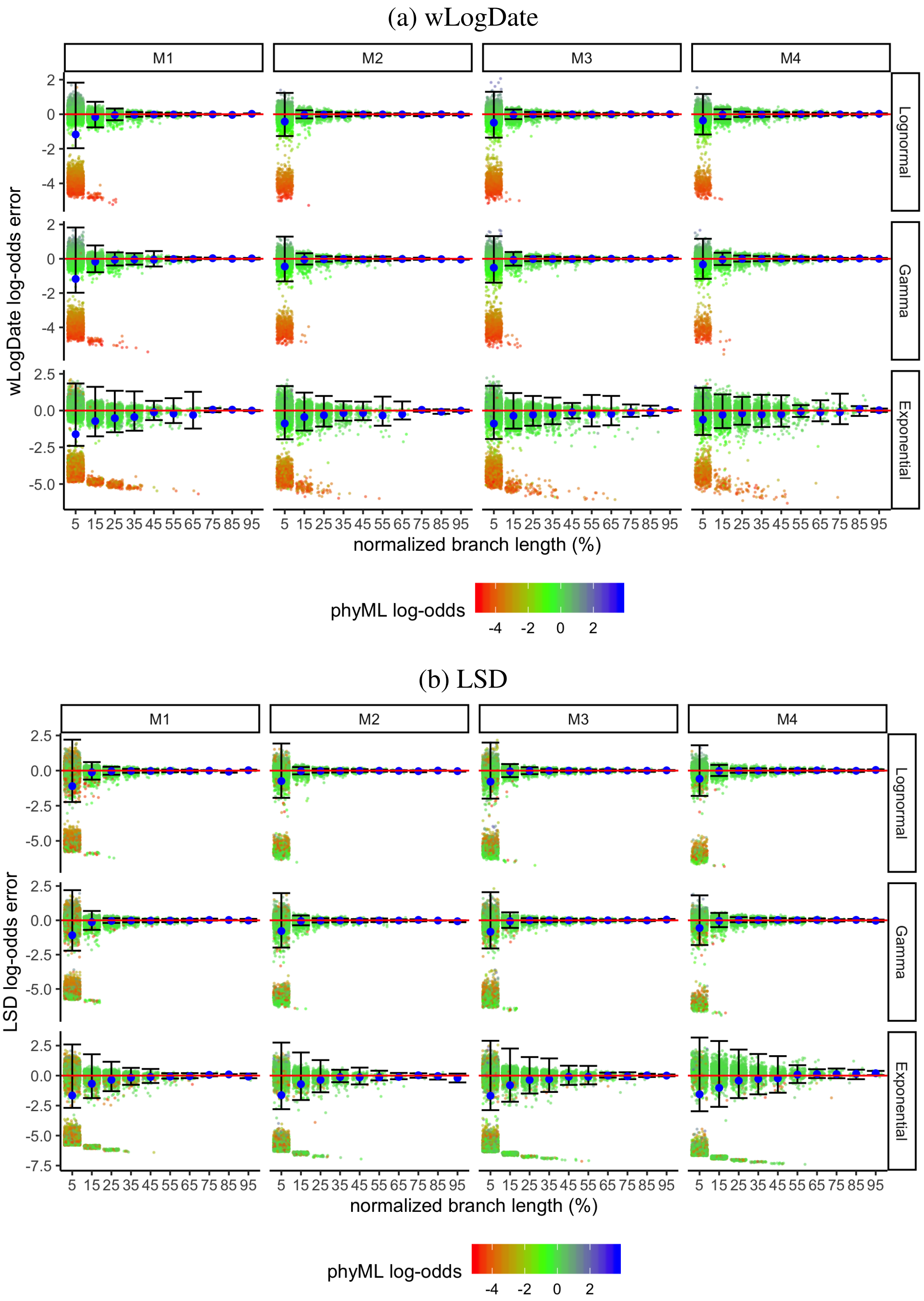


Figure D.5: A TimeTree of San Diego HIV epidemic according to wLogDate.


(a) wLogDate

(b) LSD

Figure D.6: Effects of PhyML estimation error on wLogDate and LSD performance. Figure shows log-odds error of (a) wLogDate and (b) LSD versus true branch length (in time unit); x-axis is normalized by the maximum tree branch; dots are colored by log-odds error of phyML estimates; large blue dots show means and bars show one standard deviations around medians.

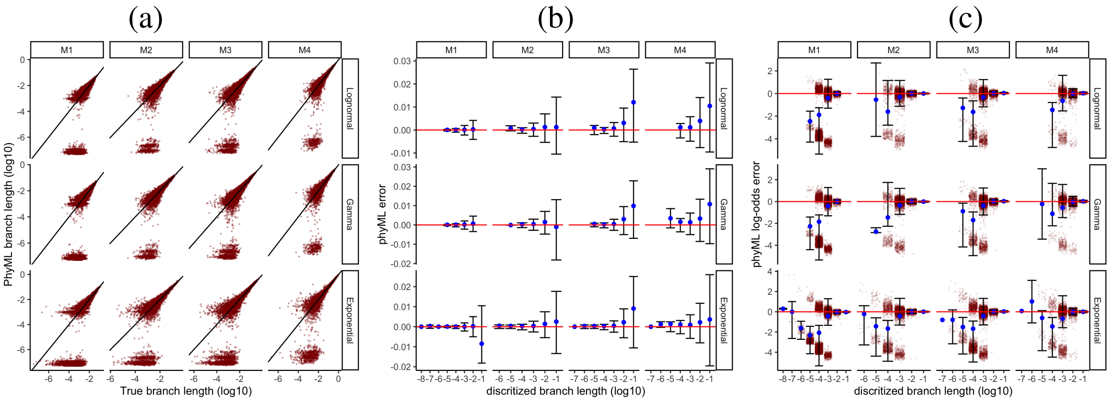


Figure D.7: Analyses of the estimated branch lengths using PhyML on simulated data. (a) Estimated versus true branch lengths (expected number of substitutions per site); axes scaled in log10. (b) Error versus true branch lengths; blue dots represent means and bars represent standard deviations around medians. (c) Log-odds error versus true branch lengths; blue dots represent means and bars represent standard deviations around medians.

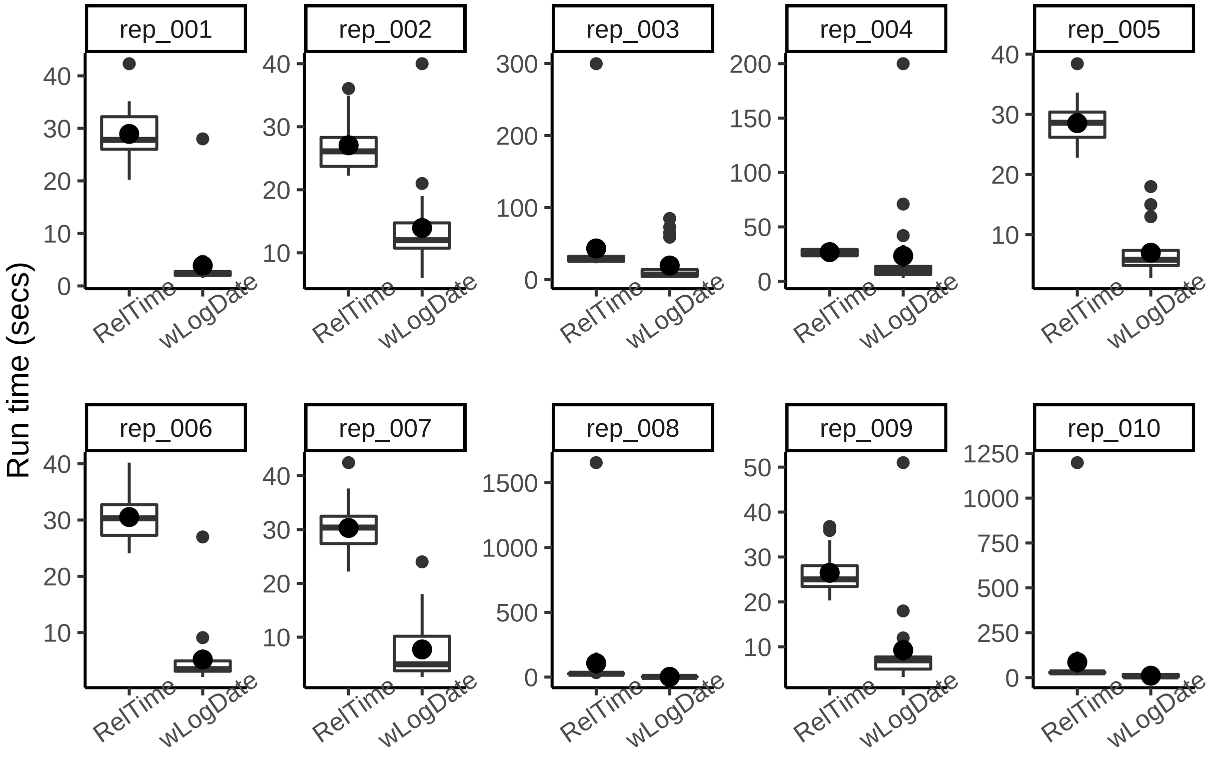


Figure D.8: Run time of wLogDate and RelTime on the 10 replicates. Box plots show distributions for the 20 tests.

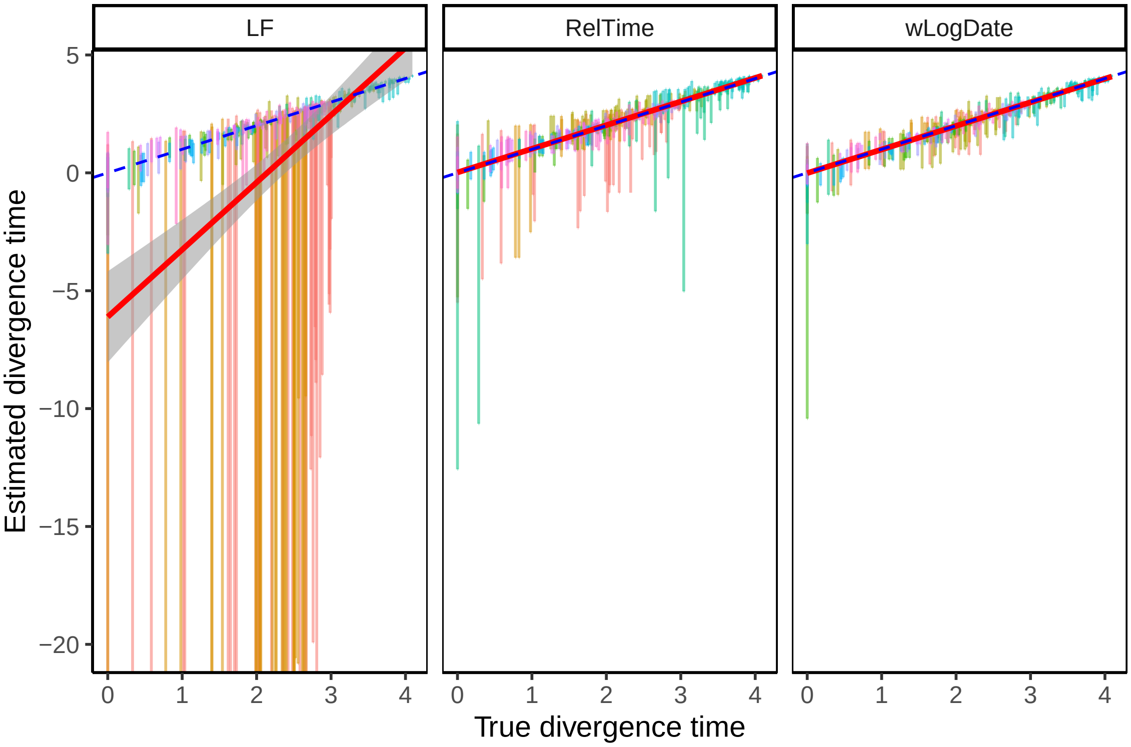


Figure D.9: Comparison of LF, RelTime, and wLogDate on the simulated data with autocorrelated rate model. The y-axis shows estimated divergence times of uncalibrated internal nodes while the x-axis shows the true divergence time. Each bar shows the 2.5% and 97.5% quantiles of the estimates of a single node's divergence time across 20 tests, each of them with different random choices of calibration points (thus, these are not CIs for one run). There are 10 replicate trees, each with 44 uncalibrated nodes (thus, 440 bars in total). There are two tests where LF produces extremely erroneous time trees (test 2 of replicate 1 and test 16 of replicate 2) and were discarded in Fig. 4.6 in the main text. The normalized RMSE of LF are 41.3 and 167.8 for these two tests, while the overall error without these two tests is 0.09. Here we show the full results for completeness. Colors are used to distinguish between replicates.

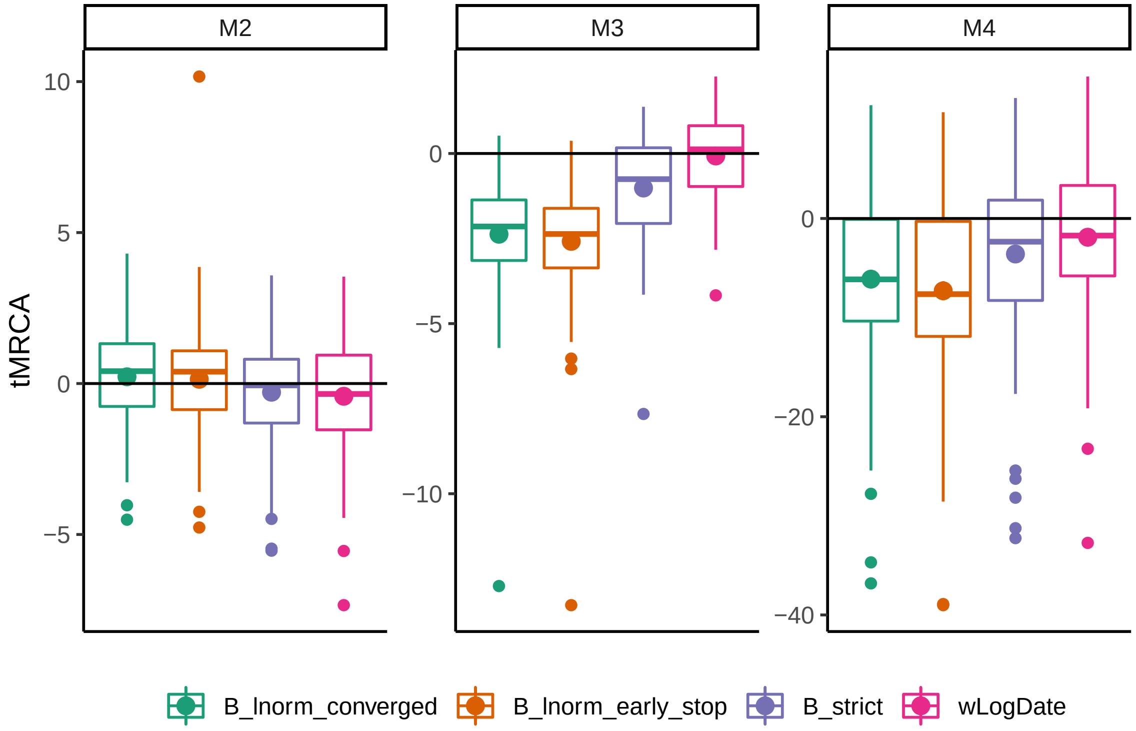


Figure D.10: Estimation of the tMRCA of *M*2, *M*3, and *M*4 of the simulated data with Lognormal clock model. For each model, BEAST was run with 3 conditions: B_strict uses the strict-clock prior, B_lnorm_early_stop uses Lognormal clock prior with MCMC chain of 10 millions, and B_lnorm_converged uses Lognormal clock prior with elongated MCMC chain to guarantee convergence (refer to table D.6 for parameters and convergence check.)

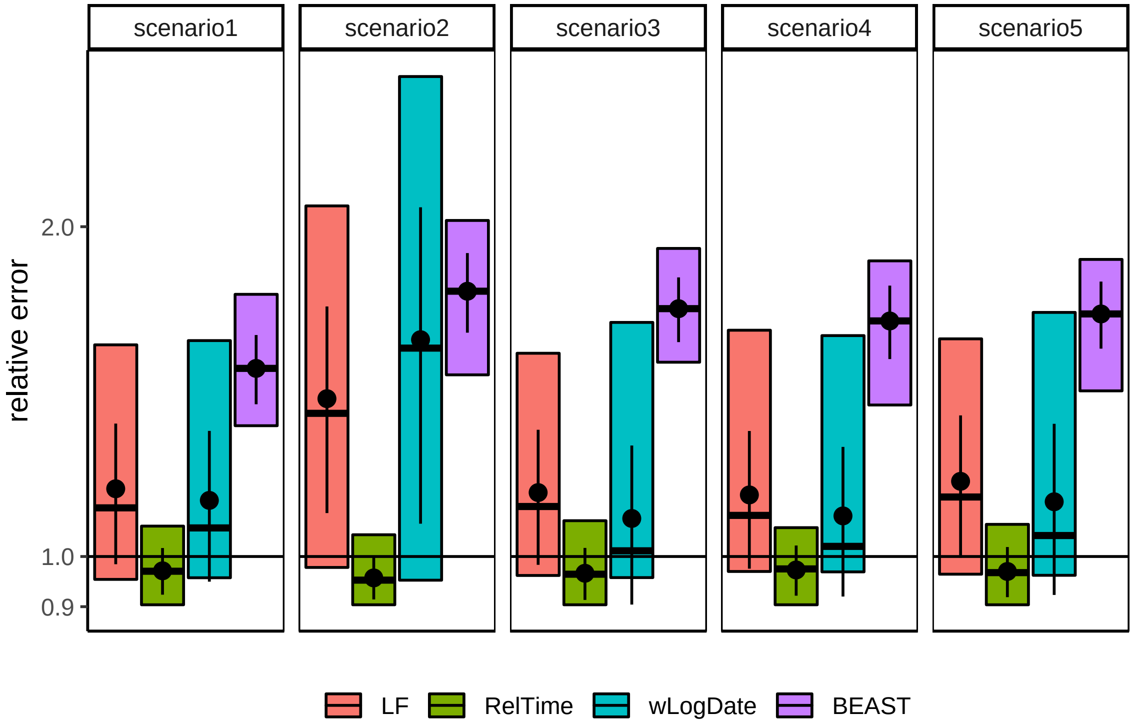


Figure D.11: Relative error of wLogDate, RelTime, and LF on inferring the Angiosperm's age on different settings of the simulation by [23]. Boxplots show median with 95% CI. Point ranges show mean with one standard deviation. Results for BEAST was obtained from the original study [23] on the same dataset but some different settings for the calibrations.

Appendix E

Supplementary materials for Chapter 5

E.1 Supplementary text

E.1.1 Solving for ω in the M-step

Here we present the algorithm to solve ω in the first round of MD-Cat where we add a constraint on ω to fix the average to a constant μ. The algorithm to solve ω in the second round of MD-Cat without this constraint is a simpler version and is not shown.

The optimization problem

Here we let τ fixed in Eq. 5.6 and solve the following optimization problem:

$$\mathcal{P} : \min_{\omega} \sum_{i=1}^{N} \sum_{j=1}^{k} \frac{q_{ij}}{\hat{b}_i} (\hat{b}_i - \omega_j \tau_i)^2 \tag{E.1}$$

such that $\omega \geq \varepsilon$ and $\sum_{j=1}^{k} \omega = k\mu$, where ε and μ are constants and $0 \leq \varepsilon < \mu$. It is trivial to see that the optimization problem $\mathcal{P}$ is convex. Therefore, we can find the global optimal ω^* of $\mathcal{P}$ using the active-set method [235]. In addition, thanks to the simple weighted least-square form of the objective function, we can solve *each iteration* of the active-set method in $O(k)$, as shown below.

Overview of the active-set method

Recall that a *feasible point* of an optimization problem is a point that satisfies all the problem's constraints. If $\omega^{(i)}$ is a feasible point of $\mathcal{P}$, then its *active-set* $\mathcal{A}^{(i)}$ is defined as:

$$\mathcal{A}^{(i)} = \{j|\omega_j^{(i)} = \varepsilon\} \tag{E.2}$$

Starting with a feasible point $\omega^{(1)}$ and its active-set $\mathcal{A}^{(1)}$, the active-set algorithm repeat the following procedure in each iteration (i) until the optimal point is found:

- Solve the equality constrained problem $\mathcal{P}^{(i)}$ defined by the active-set $\mathcal{A}^{(i)}$ (the formal definition of $\mathcal{P}^{(i)}$ will be shown later).
- Use $\omega^{(i)}$ and the optimal point $\omega^{*(i)}$ of $\mathcal{P}^{(i)}$ to find a new *feasible* point $\omega^{(i+1)}$ that is closer to the optimal point of $\mathcal{P}$ than $\omega^{(i)}$.
- Compute the active-set $\mathcal{A}^{(i+1)}$ of $\omega^{(i+1)}$.
- Replace $\omega^{(i)}$ with $\omega^{(i+1)}$ and $\mathcal{A}^{(i)}$ with $\mathcal{A}^{(i+1)}$, then repeat the procedure.

The subproblem $\mathcal{P}^{(i)}$ and the Lagrange method

In iteration (i) of the active-set method, we define and solve the following equality constrained optimization problem:

$$\mathcal{P}^{(i)} : \min_{\omega} \sum_{i=1}^{N} \sum_{j=1}^{k} \frac{q_{ij}}{\hat{b}_i} (\hat{b}_i - \omega_j \tau_i)^2 \tag{E.3}$$

such that $\omega_j = \varepsilon, \forall j \in \mathcal{A}^{(i)}$ and $\sum_{j=1}^{k} \omega = k\mu$.

$\mathcal{P}^{(i)}$ can be solved *analytically* by introducing Lagrange multipliers:

$$L = \sum_{i=1}^{N}\sum_{j=1}^{k} \frac{q_{ij}}{\hat{b}_i}(\hat{b}_i - \omega_j\tau_i)^2 - \eta\left(\sum_{j=1}^{k}\omega_j - k\mu\right) - \sum_{j\in\mathcal{A}^{(i)}} \lambda_j(\omega_j - \varepsilon), \tag{E.4}$$

where η is the Lagrange multiplier of the constraint $\sum_{j=1}^{k}\omega = k\mu$ and each λ_j is the Lagrange multiplier of the constraint $\omega_j = \varepsilon$. We will use λ to represent the vector containing all these Lagrange multipliers. We have:

$$\frac{\partial L}{\partial \omega_j} = \sum_{i=1}^{N}\left(\frac{2q_{ij}\tau_i}{\hat{b}_i}(\omega_j\tau_i - \hat{b}_i)\right) - \eta - \lambda_j \mathrm{I}_{j\in\mathcal{A}^{(i)}}, \tag{E.5}$$

where I is the indicator function.

Let ω^*, η^*, and λ^* denote the critical point of L. If $j \in \mathcal{A}^{(i)}$, then $\omega^*_j = \varepsilon$. Otherwise, ω^*_j can be solved by setting $\frac{\partial L}{\partial \omega_j}$ to 0. Thus, we have

$$\omega^*_j = \begin{cases} \varepsilon & j \in \mathcal{A}^{(i)} \\ \frac{a_j}{c_j} + \frac{\eta^*}{kc_j} & j \notin \mathcal{A}^{(i)}, \end{cases} \tag{E.6}$$

where $a_j = 2\sum_{i=1}^{N} q_{ij}\tau_i$ and $c_j = 2\sum_{i=1}^{N}\frac{q_{ij}\tau_i^2}{\hat{b}_i}$. Substitute this equation to the constraint $\sum_{j=1}^{k}\omega^*_j = k\mu$ and solve for η^*, we have:

$$\eta^* = \frac{k^2\mu - k\sum_{j=1}^{k}\frac{a_j}{c_j} - k\varepsilon|\mathcal{A}^{(i)}|}{\sum_{j=1}^{k}\frac{1}{c_j}}, \tag{E.7}$$

where $|\mathcal{A}^{(i)}|$ denotes the cardinality of $\mathcal{A}^{(i)}$. Next, substitute $\omega^*_j = \varepsilon$ to Eq. E.5 and set $\frac{\partial L}{\partial \omega_j}$ to 0, we can solve for each λ^*_j where $j \in \mathcal{A}^{(i)}$:

$$\lambda^*_j = c_j\varepsilon - \frac{\eta^*}{k} - a_j \tag{E.8}$$

Substitute η^* in Eq. E.7 to Eq. E.8 and Eq. E.6, we obtain the analytical solution to all λ_j^* and ω_j^*. Note that we can pre-compute all a_j and c_j in Eq. E.6 for all subproblems $\mathcal{P}^{(i)}$. Therefore, the analytical solution of each $\mathcal{P}^{(i)}$ can be computed in $O(k)$, as shown in Algorithm 6.

Algorithm 6 The Lagrange method to solve a subproblem $\mathcal{P}^{(i)}$ defined by the active set $\mathcal{A}^{(i)}$

```
  function SolveLagrange($\mathcal{A}^{(i)}$)
      $\omega \leftarrow []$                                        ▷ initialize to an empty list
      $\lambda \leftarrow \{\}$                                     ▷ initialize to an empty dictionary
      compute $\eta$ using Eq. E.7
      for $j \in [k]$ do
          if $j \in \mathcal{A}^{(i)}$ then
              compute $\lambda_j$ using Eq. E.8
              $\lambda[j] \leftarrow \lambda_j$
              $\omega_j \leftarrow \varepsilon$
          else:
              compute $\omega_j$ using Eq. E.6
          append $\omega_j$ to $\omega$
      return $\omega, \eta, \lambda$
```

Computing $\omega^{(i+1)}$ and $\mathcal{A}^{(i+1)}$

Let $\omega^{(i)}$ and $\mathcal{A}^{(i)}$ denote the feasible point and its active-set found in iteration (i) of the active-set algorithm and let $\omega^{*(i)}$ denote the optimal point of $\mathcal{P}^{(i)}$, η^* and $\lambda^{*(i)}$ denote its optimal Lagrange multipliers. Depending on the characteristics of $\omega^{*(i)}$ and $\lambda^{*(i)}$, we can find $\omega^{(i)}$ and $\mathcal{A}^{(i)}$ as follows.

Case 1: $\omega^{*(i)}$ is feasible to $\mathcal{P}$ and $\lambda^{*(i)} \geq 0$ In this case, $\omega^{*(i)}$ is also the optimal point of $\mathcal{P}$, according to the KKT condition. We simple return $\omega^{*(i)}$.

Case 2: $\omega^{*(i)}$ is feasible to $\mathcal{P}$ and $\exists \lambda_j^{*(i)} < 0$ In this case, set $\omega^{*(i+1)}$ to $\omega^{*(i)}$ and find the constraint j that has the most negative $\lambda_j^{*(i)}$ and remove it from $\mathcal{A}^{(i)}$ to get $\mathcal{A}^{(i+1)}$ (i.e. relax the "useless" constraint).

Case 3: $\omega^{*(i)}$ is infeasible to $\mathcal{P}$ In this case, we search for a new feasible point $\omega^{(i+1)}$ that is closer to the optimum and update the active-set. To this purpose, we start from the previous feasible point $\omega^{(i)}$ and move it as close as possible to $\omega^{*(i)}$ on the direction $d = \omega^{*(i)} - \omega^{(i)}$ such that the new point is still feasible. In other words, we need to find the largest number $\alpha \in [0,1]$ such that

$$\omega_j^{(i)} + \alpha(\omega_j^{*(i)} - \omega_j^{(i)}) \geq \varepsilon, \forall j \in \mathcal{A}^{(i)} \tag{E.9}$$

Let $V = \{j | \omega_j^{*(i)} < \varepsilon\}$ denote the *violated set* of $\omega^{*(i)}$. Because $\omega^{(i)}$ is feasible, it is easy to see that Eq. E.9 is always satisfied for all $j \notin V$. Thus, to find α we only need to satisfy Eq. E.9 for all j in the violated set V. Now we have two sub-cases:

- If there exists $j \in V$ such that $\omega_j^{(i)} = \varepsilon$, then Eq. E.9 is satisfied only if $\alpha(\omega_j^{*(i)} - \omega_j^{(i)}) \geq 0$. On the other hand, because $j \in V$ and $\omega^{(i)}$ is feasible, we have $\omega_j^{*(i)} < \varepsilon \leq \omega_j^{(i)} \implies \omega_j^{*(i)} - \omega_j^{(i)} < 0$. Therefore, Eq. E.9 is satisfied only if $\alpha = 0$. Thus, in this case we set $\omega^{(i+1)} = \omega^{(i)}$ and add the constraint j into $\mathcal{A}^{(i)}$ to obtain $\mathcal{A}^{(i+1)}$.

- Otherwise, let $\Delta_j = \frac{\omega_j^{*(i)} - \varepsilon}{\omega_j^{(i)} - \varepsilon}$ for all $j \in V$. After rewriting Eq. E.9 and substituting Δ_j to it, we get the following condition: $\alpha \leq \frac{1}{1-\Delta_j}$ for all $j \in V$, or equivalently, $\alpha \leq \min_{j \in V} \frac{1}{1-\Delta_j} = \frac{1}{1-\Delta_p}$ where Δ_p is the minimum of all Δ_j. Thus, we set $\alpha = \frac{1}{1-\Delta_p}$, $\omega^{(i+1)} = \omega^{(i)} + \alpha d$, and add p into $\mathcal{A}^{(i)}$ to obtain $\mathcal{A}^{(i+1)}$.

The active-set algorithm described in this section is summarized in Algorithm 7.

E.1.2 Solving for τ in the M-step

Here we let ω fixed in Eq. 5.6 and solve the following optimization problem:

$$\mathcal{P} : \min_{\tau} \sum_{i=1}^{N} \sum_{j=1}^{k} \frac{q_{ij}}{\hat{b}_i} (\hat{b}_i - \omega_j \tau_i)^2 \tag{E.10}$$

Algorithm 7 The active-set method to solve the problem $\mathcal{P}$ defined in Eq. E.1

```
function SolveOmega($\tau, \hat{b}, \mu, \varepsilon, k$)
    $\omega_j^{(1)} \leftarrow \mu$ for all $j = 1..k$                                   ▷ feasible initital point
    $\mathcal{A}^{(1)} \leftarrow \emptyset$
    for i = 1 to MaxNumberIterations do
        $\omega^{*(i)}, \eta^{*(i)}, \lambda^{*(i)} \leftarrow$ SOLVELAGRANGE($\mathcal{A}^{(i)}$)                  ▷ See Algorithm. 6
        if $\omega^{*(i)}$ is feasible then
            if $\lambda_j^{*(i)} \geq 0$ for every $j \in \mathcal{A}^{(i)}$ then
                return $\omega^{*(i)}$                    ▷ satisfies KKT $\Longrightarrow$ feasible and optimal
            else
                $\lambda_h^{*(i)} \leftarrow \min_{j \in \mathcal{A}^{(i)}} \lambda_j^{*(i)}$
                remove $h$ from $\mathcal{A}^{(i)}$ to get $\mathcal{A}^{(i+1)}$
        else
            $V = \{j | \omega_j^{*(i)} < \varepsilon\}$                                ▷ the violated set of $\omega^{*(i)}$
            if there exists $j \in V$ s.t. $\omega_j^{(i)} = \varepsilon$ then
                add $j$ into $\mathcal{A}^{(i)}$ to get $\mathcal{A}^{(i+1)}$
                $\omega^{(i+1)} \leftarrow \omega^{(i)}$
            else
                $\Delta_j \leftarrow \frac{\omega_j^{*(i)} - \varepsilon}{\omega_j^{(i)} - \varepsilon}$ for all $j \in V$
                $\Delta_p \leftarrow \min_{j \in V} \Delta_j$
                $\alpha \leftarrow \frac{1}{1 - \Delta_p}$
                $\omega^{(i+1)} \leftarrow \omega^{(i)} + \alpha(\omega^{*(i)} - \omega^{(i)})$              ▷ feasible and “more optimal”
                add $p$ into $\mathcal{A}^{(i)}$ to get $\mathcal{A}^{(i+1)}$
    return the last $\omega^{(i)}$
```

such that $\tau_i \geq \varepsilon$, where ε is a non-negative constant, and the linear constraints Ψ set by the calibration points are satisfied. It is trivial to see that $\mathcal{P}$ is convex, so we can find the global optimal τ^* using active-set method. Furthermore, thanks to the simple weighted least-square form of the objective function and the hierarchical structure of the linear constraints, we can solve each iteration of the active-set method in $O(N)$. Below we derive the analytical solution for each iteration of the active-set method. Readers can refer to [235] and section E.1.1 of this text to derive the full active-set procedure.

Notations

In this section, we use the following extended notations.

- $P(x,y)$: the path between the two nodes x and y on the tree. The nodes are uniquely identified by their indices.
- $P(x)$: the path from the root to node x.
- n_i: the number of nodes below node i.
- $\mathcal{C}_i$: the set of calibration points below node i.

The linear constraints

In our algorithm, we use the original setup of the linear constraints Ψ, where the unknown divergence time t_0 at the root is added to the equations and is co-estimated with τ. The p calibration points $t_1,\ldots,t_p$ define a set of p constraints $C_1,...,C_p$:

$$C_i : \sum_{j \in P(i)} \tau_j = t_i - t_0 \tag{E.11}$$

The subproblem $\mathcal{P}^{(h)}$ and the Lagrange method

In iteration (h) of the active-set method, we define and solve the following equality constrained optimization problem:

$$\mathcal{P}^{(h)} : \min_{\tau} \sum_{i=1}^{N}\sum_{j=1}^{k} \frac{q_{ij}}{\hat{b}_i}(\hat{b}_i - \omega_j\tau_i)^2 \tag{E.12}$$

such that $\tau_i = \varepsilon, \forall i \in \mathcal{A}^{(h)}$, where $\mathcal{A}^{(h)}$ is the active-set at iteration (h), and all the p linear constraints $C_1, C_2, ..., C_p$ are satisfied. Introducing Lagrange multipliers, we have:

$$L = \sum_{i=1}^{N}\sum_{j=1}^{k} \frac{q_{ij}}{\hat{b}_i}(\hat{b}_i - \omega_j\tau_i)^2 - \sum_{m=1}^{p}\lambda_m\left(t_0 + \sum_{i\in P(m)}\tau_i - t_m\right) - \sum_{i\in\mathcal{A}^{(h)}}\nu_i(\tau_i - \varepsilon), \tag{E.13}$$

where λ_m and ν_i are Lagrange multipliers. Taking partial derivatives at each τ_i and t_0, we have:

$$\frac{\partial L}{\partial \tau_i} = \alpha_i\tau_i + \beta_i - \sum_{m\in\mathcal{C}_i}\lambda_m - \nu_i \mathrm{I}_{i\in\mathcal{A}^{(h)}}, \tag{E.14}$$

where $\alpha = \sum_{j=1}^{k}\frac{2q_{ij}\omega_j^2}{\hat{b}_i}$, $\beta = -\sum_{j=1}^{k}2q_{ij}\omega_j$, and $\mathcal{C}_i$ is the set of calibration points below node i;

$$\frac{\partial L}{\partial t_0} = -\sum_{m=1}^{p}\lambda_m \tag{E.15}$$

Setting the partial derivatives to 0 and use the constraints in $\mathcal{A}^{(h)}$, we have the following system of equations:

$$\tau_i^* = \begin{cases} \varepsilon & i \in \mathcal{A}^{(h)} \\ \frac{1}{\alpha_i}\sum_{m\in\mathcal{C}_i}\lambda_m - \frac{\beta_i}{\alpha_i} & i \notin \mathcal{A}^{(h)} \end{cases} \tag{E.16}$$

$$\nu_i^* = \frac{-1}{\alpha_i}\sum_{m\in\mathcal{C}_i}\lambda_m + \frac{\beta_i}{\alpha_i} + \varepsilon \tag{E.17}$$

$$\sum_{m=1}^{p} \lambda_m^* = 0 \tag{E.18}$$

Note that using Eq. E.16, we only need the Lagrange multipliers *below* τ_i^* to compute it. Therefore, in a postorder traversal (with complexity $O(N)$), it is straight-forward to see that we can use the method of substitution to compute all τ_i^* and λ_m^* with respect to λ_1^* (or any other arbitrarily chosen λ_j^*) using the method of substitution (a similar strategy has been described in [340]). Thus, all λ_j^* can be parameterized by λ_1^* as follow:

$$\lambda_m^* = \gamma_{m,1}\lambda_1^* + \eta_{m,1}, \tag{E.19}$$

where $\gamma_{m,1}$ and $\eta_{m,1}$ are constants. Substituting to Eq. E.18, we obtain the analytical solution for λ_1^*:

$$\lambda_1^* = \frac{-\sum_{m=2}^{p} \eta_{m,1}}{1 + \sum_{m=2}^{p} \gamma_{m,1}} \tag{E.20}$$

Substituting back to Eqs. E.16, E.17, and E.19, we obtain the analytical solution for all τ_i^*, λ_m^*, and ν_i^*.

E.1.3 Convergence of the EM algorithm

Recall that in the M-step, we need to find $\tau \geq 0$ and $\omega \geq \varepsilon$ that satisfy Ψ and maximize the following function:

$$g(\tau, \omega; q) = \textstyle\sum_{i=1}^{N} \sum_{j=1}^{k} q_{ij} \log f(\hat{b}_i | \omega_j, \tau_i), \tag{E.21}$$

where f denote the density function of the Gaussian model for branch estimation uncertainty, as described in the main text.

At each iteration (h) of the algorithm, let $q^{(h)}$ denote the posterior computed in the E-step, $\tau(h), \omega(h)$ and $\tau(h)+1, \omega(h+1)$ denote the solution found before and after M-step. Then we claim the following:

Claim 8. If $g(\tau^{(h+1)},\omega^{(h+1)};q^{(h)}) \geq g(\tau^{(h)},\omega^{(h)};q^{(h)})$, then $l(\tau^{(h+1)},\omega^{(h+1)}) \geq l(\tau^{(h)},\omega^{(h)})$

Proof. Let $C = \sum_{i,j} q_{ij} \log(kq_{ij}^{(h)})$. Subtracting C from both sides of the conditioned inequality, we have:

$$\sum_{i=1}^{N}\sum_{j=1}^{k} q_{ij}^{(h)} \log \frac{f(\hat{b}_i|\omega_j^{(h+1)},\tau_i^{(h+1)})}{kq_{ij}^{(h)}} \geq \sum_{i=1}^{N}\sum_{j=1}^{k} q_{ij}^{(h)} \log \frac{f(\hat{b}_i|\omega_j^{(h)},\tau_i^{(h)})}{kq_{ij}^{(h)}} \tag{E.22}$$

Recall that $q_{ij}^{(h)} = \frac{f(\hat{b}_i|\omega_j^{(h)},\tau_i^{(h)})}{\sum_{m=1}^{k} f(\hat{b}_i|\omega_m^{(h)},\tau_i^{(h)})}$ and $\sum_j q_{ij}^{(h)} = 1$. Therefore, we can rewrite the right hand side (RHS) of Eq. E.22 as follows

$$\begin{aligned}
\text{RHS} &= \textstyle\sum_{i=1}^{N}\sum_{j=1}^{k} q_{ij}^{(h)} \log\left[\frac{1}{k}\sum_{m=1}^{k} f(\hat{b}_i|\omega_m^{(h)},\tau_i^{(h)}\right] \\
&= \textstyle\sum_{i=1}^{N} \log\left[\frac{1}{k}\sum_{m=1}^{k} f(\hat{b}_i|\omega_m^{(h)},\tau_i^{(h)}\right]\sum_{j=1}^{k} q_{ij}^{(h)} \\
&= \textstyle\sum_{i=1}^{N} \log\left[\frac{1}{k}\sum_{m=1}^{k} f(\hat{b}_i|\omega_m^{(h)},\tau_i^{(h)})\right] = l(\tau^{(h)},\omega^{(h)})
\end{aligned} \tag{E.23}$$

On the other hand, applying Jensen's inequality, we get an upper bound for the left hand side (LHS) of Eq. E.22:

$$\begin{aligned}
\text{LHS} &\leq \textstyle\sum_{i=1}^{N} \log\left[\sum_{j=1}^{k} q_{ij}^{(h)} \frac{f(\hat{b}_i|\omega_j^{(h+1)},\tau_i^{(h+1)})}{kq_{ij}^{(h)}}\right] \\
&= \textstyle\sum_{i=1}^{N} \log\left[\frac{1}{k}\sum_{j=1}^{k} f(\hat{b}_i|\omega_j^{(h+1)},\tau_i^{(h+1)})\right] = l(\tau^{(h+1)},\omega^{(h+1)})
\end{aligned} \tag{E.24}$$

Thus, from Eq. E.22, Eq. E.23, and Eq. E.24, we have $l(\tau^{(h+1)},\omega^{(h+1)}) \geq l(\tau^{(h)},\omega^{(h)})$. □

Corollary 4. *The EM algorithm described in the main text monotonically improves the log-likelihood function after each iteration and converges to a local minimum.*

Proof. Recall that in the M-step at iteration (h) , we use coordinate descent starting at $(\tau^{(h)},\omega^{(h)})$ to find $(\tau^{(h+1)},\omega^{(h+1)})$. Therefore, $g(\tau^{(h+1)},\omega^{(h+1)};q^{(h)}) \geq g(\tau^{(h)},\omega^{(h)};q^{(h)})$ by construction. By Claim 8, we conclude that the algorithm monotonically improves l. In addition, because the

Gaussian PDF is bounded above, l is also bounded above. Thus, the sequence $sn(h) = l(\tau^{(h)}, \omega^{(h)})$ is monotonically increasing and bounded above, so it converges. $\square$

Appendix F

Supplementary materials for Chapter 6

F.1 Supplementary Notes

F.1.1 Novel algorithms for prototype selection

Given a distance matrix D for n objects and a given number k, the problem of *prototype selection* is to find a subset of $k \subset n$ objects, with $1 < k < n$, such that an objective function d is optimized. This problem is known to be NP-hard [104]. In the example of [57], the objects are geographical locations of n clients of a banking corporation. The distance matrix D reflects the time to clear a check drawn in client's location i and cashed in client's location j. The bank's problem is to decide for a given number k at what client locations to open a branch in order to maximize their available funds. Thus, the objective function is the minimization over the given distances in D. For our use case of choosing a most representative subset of k genomes, we maximize over the given distance matrix as defined by MinHash signatures [240] in order to maximize diversity. An exact algorithm must enumerate all n over k combinations of k objects, compute the score for every combination via objective function d and select optimal combination(s). Since n over k grows exponentially, this is impractical for relevant input sizes and we have to resort to heuristics. Fortunately, results of alternative heuristics implementations can be compared by their score, although it remains unclear what an optimal score would be.

We devised a naive algorithm to heuristically solve the prototype selection problem: It starts with the full set of n objects. The initial score for all n objects is the sum of pairwise distances for all objects in n. In each iteration, we greedily choose the one object, which reduces the overall score the least and remove it from the shrinking set. We continue until k is reached. We call this algorithm due its shrinking nature of maximizing overall distance score: "**destructive_maxdist**". We furthermore implemented alternative algorithms to solve the prototype selection problem. The implementation "**constructive_maxdist**" is a close relative: We start with the two objects that are most distant from each other in D. The set of prototypes is then constructively grown by adding the object showing largest sum of distances to all remaining objects in D. The method "**constructive_protoclass**" implements the algorithm of [26] but for only one "class". Intuitively, a sphere is drawn around every object in D with radius ε. The element whose ball covers most other objects is selected as prototype. All such covered elements and the new prototype are removed for the next round. This is repeated until no balls cover more than its center element. Our fourth and last method "**constructive_pMedian**" implements the *p-median* algorithm of [205] which is closely related to k-means clustering for given k.

Our comparison of those four implementations of heuristic algorithms to solve the prototype selection problem shows that "destructive_maxdist" requires least run time, returns highest scores for many instances and can handle instances of $n = 90,000$ within seconds (Supplementary Fig. F.1b-d).

For our application, we needed to extend the original problem definition by allowing the pre-definition of r objects as prototypes, a.k.a., "*seeds*". Thus, $k - r$ prototypes need to be selected from n objects such that all objects of r are guaranteed to become prototypes. This alternation will preserve objects of biological interest while minimizing the reduction of score. For example, we wanted to make sure that several well-studied *E. coli* strains are chosen over other thousands of less popular ones. The algorithms work as described above, but in the initiation phase, the set of selected prototypes is not empty but filled with *r* objects and corresponding rows and columns in *D* are masked. The increase in runtime is marginal with this function enabled, while the resulting score is notably higher than that by not using this function (Supplementary Fig. F.1e).

Python implementations are provided at https://github.com/biocore/wol/, under directory code/prototypeSelection, which also contains a Jupyter notebook we used for benchmarking.

F.1.2 Comparative analysis of trees by different methods and input data

We conducted a systematic exploration of the optimal strategy for building the microbial tree of life. Multiple species trees were reconstructed, using differential taxon, gene and site sampling strategies, as well as different tree-building methods, implementations and evolution models. The comparative analysis results are detailed in this section. Two metrics were mainly used for comparing trees: 1) The Robinson-Foulds (RF) distance [277] normalized by tree sizes, which measures the topological discrepancy between each pair of trees; 2) "tip distance" (TT), which measures the correlation between tip-to-tip distance matrices of two trees (see Methods). In addition, the distributions of branch support values, if comparable and relevant, were addressed. To maximize objectiveness, these analyses are purely based on the mathematical properties of trees and are free from any biological knowledge.

Comparison between "full-scale" ASTRAL and CONCAT trees. The two tree-building methods produced similar species tree topologies (Fig. 6.3). The distance between the two CONCAT trees is shorter (RF=0.179) than between either of them and the ASTRAL tree (see below), which is expected considering the differential mechanisms behind each method. The CONCAT tree based on randomly sampled sites ("concat.rand") resembles the ASTRAL tree (RF=0.260) more than does the CONCAT tree based on most conserved sites ("concat.cons") (RF=0.312), likely a consequence of random site sampling, which better represents the full-length sequence alignments that were used for building individual gene trees for ASTRAL. Additionally, the species tree built on all sites but using FastTree ("fasttree") shows higher similarity with the random CONCAT tree (RF=0.156) than with the ASTRAL tree (RF=0.257), implicating higher impact by tree-building method than by robustness of the same method (further discussed below, see Supplementary Fig. F.12). Interestingly, the tree based on ribosomal proteins ("concat.rpls") is more similar with the ASTRAL tree (RF=0.253) than with the CONCAT trees (RF=0.340 (conserved) / 0.304 (random)).

The two methods for building large phylogenies have different computational requirements (Supplementary Table F.2). As mentioned before, computer memory and run-time constraints limit the size of the datasets and the complexity of the models that can be analyzed with CONCAT. On the other hand, the gene tree summary method implemented in ASTRAL is less constrained, even though its overall cost is greater, because most of the time is spent in building individual gene trees, a step that can be fully parallelized across compute nodes. This scalability of ASTRAL means that it can be extended in a straightforward manner to even larger scale phylogenomic analyses than considered here. We estimated branch lengths for the ASTRAL tree using either most conserved or randomly selected sites (see Methods). Even though random site sampling gave a larger tree dimension overall than conserved site sampling, the individual branch lengths had strong linear correlation between the two methods (slope=1.776, $R^2 = 0.974, p = 0.0$).

Evaluation of trees inferred using implicit vs. explicit methods. We tested three alternative approaches for assessing the relationships among organisms: namely, either explicit (gene tree summary or gene alignment concatenation) or implicit (by marker gene distribution, MinHash signature, andor NCBI taxonomy. Albeit simple and applicable approaches, they do not explicitly model the evolutionary process of molecular sequences. The topological distances among these trees and the species trees reconstructed using dedicated phylogenetic approaches are shown in Supplementary Fig. F.7. It reveals high discrepancy among the three implicit trees and from the explicit category (RF > 0.62). In particular, the taxonomy has the highest discrepancy (RF > 0.83), due to its over-simplified hierarchies. Meanwhile, the four phylogenetic trees, despite using different gene selection, site sampling and tree-building methods, notably converge better (RF < 0.35). Topologies were compared using the Robinson–Foulds (RF) metric [277]. The topologies of the trees built using explicit methods, either summary or concatenation, are better converged than those obtained from the alternate, cheaper methods, which do not directly operate on sequence data (Supplementary Fig. F.7a). This underscores the necessity of using sequence data and dedicated phylogenetic approaches to accurately define evolutionary relationships in high-quality phylogenomic studies.

Impact of gene tree quality and quantity on ASTRAL trees. We evaluated whether a large number of loci, i.e., the practice of "phylogenomics" is essential in resolving species evolution using ASTRAL, which is based on the summary of multiple gene tree topologies. The 381 marker genes were randomly downsampled to smaller sets, on each of which an ASTRAL tree was built. We observed a slightly increased level of deviation from the original, full-scale ASTRAL tree (Supplementary Fig. F.10a). With 200 gene trees (around half of the original 381), the topology differed by RF=0.081. Meanwhile, the branch supports (local posterior probabilities) continued to increase with the number of gene trees (Supplementary Fig. F.10c) and did not plateau even with 381 gene trees, suggesting the benefit of including more loci in resolving species phylogeny.

We also assessed the influence of gene tree quality on the ASTRAL tree. Four trees were generated for each marker gene: one by FastTree, and the other three by RAxML, either based on the FastTree starting tree or two random seeds (see Methods). The reference ASTRAL tree was built using the best scoring RAxML gene tree of the three. As alternatives, we built two more ASTRAL trees, either based on the FastTree-started RAxML trees, or the initial FastTree trees. We observed low levels of topological discrepancy from the reference ASTRAL tree (RF=0.048 and 0.090, respectively) (Supplementary Fig. F.10b) and very close branch support distributions (Supplementary Fig. F.10d).

Impact of taxon sampling on species phylogeny. A long-standing dilemma for phylogeneticists is to balance among the number of taxa, the number of sites, and the robustness of algorithm, subject to realistic computational limitations. Fewer taxa allow the use of more expensive methods (further discussed below and in the main text, also see Supplementary Fig. F.23), at the cost of losing signals that would otherwise be helpful in better defining the evolutionary relationships among clades. To test the impact of reduced taxon sampling on the species tree, we downsampled from the original 10,575 genomes to a series of fewer taxa, in each case maximizing the representativeness of the deep phylogeny of bacterial and archaeal evolution (see Methods). The three robust phylogenetic methods—ASTRAL, CONCAT conserved, and CONCAT random (which produced Supplementary Figs. F.4-F.6, respectively)—were applied to each taxon set.

As the taxon number decreased, the reconstructed topology gradually deviated from that of the full tree (Supplementary Fig. F.13a, first row of each panel). This trend was more obvious in the CONCAT trees (conserved: RF=0.138 to 0.551, random: RF=0.110 to 0.384) than in the ASTRAL trees (RF=0.056 to 0.296) (Supplementary Fig. F.13a, comparing among panels), suggesting that ASTRAL produced more stable topologies with taxon downsampling. Meanwhile, the deviation among trees by the three methods increased as the taxon number decreased (sum of RF=0.752 to 1.653) (Supplementary Fig. F.13b). These results suggest that taxon sampling does have an impact on the tree topology. Although ASTRAL appears to be more resistant to this effect than CONCAT, it still suffered with an RF=0.103 (which translates into 10.3% incongruent clades) when the taxon number went from 10,575 down to 1,000. Therefore, the quantity of taxa is important in assessing the deep phylogeny.

Impact of site sampling and alternative models on CONCAT trees. Because of the computational expense of CONCAT with RAxML, we had to truncate the concatenated sequence alignment to at most 100 sites per marker gene (see Methods), leaving approximately 38k sites in total. Although this was more than eight times as many as the PhyloPhlAn default (on average 12 sites per gene, or 4.5k sites in total), there is a considerable loss of signals from the 192k-site full alignment. Meanwhile, the "trident" algorithm implemented in PhyloPhlAn enabled selection of the most conserved sites, compensating for the potential alignment inaccuracy in the full alignment which may be deleterious in the subsequent phylogenetic inference. To assess the influence of site sampling on the species tree, we used the trident algorithm to sequentially select 100, 50, and 25 sites per gene, plus the PhyloPhlAn default ($\sim$12), and compared the CONCAT species trees generated in each case.

Simultaneously, we evaluated the two alternative methods for modeling rate heterogeneity among sites: Gamma (classical and expensive) and CAT (a faster and less memory-intensive approximation to Gamma, which produces likelihood values than cannot be compared between analyses) [320]. (Note that the rate heterogeneity discussed here should not be confused with the more complex, profile mixture models discussed below.) Due to computational constraints, RAxML analyses were not feasible with the Gamma model on more than 25 sites per gene or with the CAT model on more than 100 sites per gene. Whereas the CONCAT trees discussed in the main text were based on 100 sites per gene with the CAT method (see Methods), here we also consider trees based on either 25 sites per gene or the default setting with the Gamma model.

We observed a pattern of sequential shift in both topology and among-taxa distances along with site sampling (Supplementary Fig. F.11). From the default setting to 100 sites per gene, there was an RF=0.308 and a TT=0.099 (which translates into a Pearson correlation coefficient of 0.802). This sequence moves toward the two trees built on randomly selected sites or all sites (the later was built using FastTree, which is further discussed below, see also Supplementary Fig. F.12). The patterns suggest that site sampling does have an impact on the phylogenetic trees. Therefore we chose to discuss both CONCAT trees using most conserved or randomly selected sites in interpreting the biology behind the trees. Furthermore, we noted that the choice of CAT vs. Gamma model had low impact on tree topology and phylogenetic distances (RF=0.040 and 0.127, TT=0.00121 and 0.0046, 25 sites per gene and default).

Impact of non-vertical evolution on species phylogeny. Conventional molecular phylogenetics analyses usually attempt to avoid loci that are prone to horizontal gene transfer (HGT), which is prevalent in the microbial world and affects a large range of genes [54, 111]. One major advantage of ASTRAL is its robustness to HGTs, allowing us to include as many as 381 gene trees to achieve optimal species tree accuracy. To validate this assumption in the context of this study, we performed a test, in which the marker genes were downsampled based on the quartet score of the corresponding gene tree—a measurement of the consistency between gene and species evolution. We selected four quartet scores thresholds: 0.5, 0.67, 0.75 and 0.8, and performed both ASTRAL and CONCAT (using conserved or random sites) species tree reconstruction on subsets of marker genes above each threshold. The results show that with fewer but presumably more "vertically evolving" genes, the ASTRAL trees retained notably more consistent topologies (smaller RF distance) than CONCAT trees did (Supplementary Fig. F.22a). When all species trees were included in one matrix, we observed close clustering of the ASTRAL trees, in contrast to the diverse distribution of CONCAT trees (Supplementary Fig. F.22b, c) (ASTRAL vs. CONCAT conserved / random, PERMANOVA pseudo-$F = 5.612/5.571, p$-value=0.009 / 0.007). These observations suggest that ASTRAL is significantly more robust against gene tree discordance compared to CONCAT.

We next checked the branch supports of the ASTRAL trees. A moderate decrease along with fewer gene trees was observed (Supplementary Fig. F.22d), despite the increased overall concordance of the remaining gene trees. Together with the discussion above (see also Supplementary Fig. F.7), this again suggests the benefit of using a large number of gene trees in an ASTRAL analysis.

Evaluation of species trees built using site heterogenerous models on 1,000 taxa. The classical site homogenerous substitution model (usually referred to as Gamma or +G) [320] has been widely used in phylogenetics studies, including most modern efforts for building the microbial tree of life (e.g., [246, 144]). It assumes that all sites are subject to the same evolutionary process, with rate heterogeneity following a Gamma distribution. However studies have shown that this simplified assumption is prone to the long branch attraction (LBA) artefacts, especially with deep phylogenetic trees where large variations of evolutionary process are likely present [178, 348]. To confirm the robustness of our findings based on the use of the Gamma model, we also built CONCAT trees using the profile mixture model C60, which was shown more robust against LBA [309], together with the posterior mean site frequency (PMSF) method implemented in IQ-TREE which enables relatively large-scale analysis with this complex model [348]. Yet this method is still notably more expensive than our reference approach, and limited our analysis to only 1,000 downsampled taxa (described above, see also Supplementary Fig. F.13). For comparison, we built additional CONCAT trees on these 1,000 taxa, using either the classical Gamma model, or the FreeRate model, which relaxes from the assumption of Gamma distribution of rates [316]. We also included the 10,575-taxon trees pruned to the 1,000 taxa for comparison.

This analysis provided an alternative and highly controlled 1,000-taxon test set to compare models (Supplementary Fig. F.23) and to re-assess a series of questions discussed above. There was a relatively stable disparity between pairs of trees by conserved and random site sampling, in both topology (RF=0.201 ± 0.011, mean and std. dev., same below) and phylogenetic distances (TT=0.0439 ± 0.0022), with PMSF not being exceptional (Supplementary Fig. F.23a). Similarly, there was a relatively stable (but more variable than between site sampling) disparity between trees by 10,575 or 1,000 taxa, both built using the Gamma model (RF=0.268 ± 0.041, TT=0.0173 ± 0.0100), with random site sampling being most consistent (Supplementary Fig. F.23b). These observations largely support the findings discussion above (see also Supplementary Figs. F.11 and F.13). Interestingly, the topological inconsistency introduced by differential taxon sampling is significantly higher than by site sampling (two-tailed **t**-test **p**=0.0204), but the inconsistency in phylogenetic distances is the opposite (two-tailed *t*-test *p*=0.00198). The variance between the 381 global markers vs. the 30 ribosomal proteins was also stable, and the most significant, especially in phylogenetic distances (RF=0.372 ± 0.018, TT=0.162 ± 0.028) (Supplementary Fig. F.23d). In both PCoAs, the differential choice of loci dominated the variances on axis 1 (which explains 46.70% and 92.88% variance, respectively) (Supplementary Fig. F.23e, f).

Now consider differential site heterogeneity models: For each dataset, trees generated by the Gamma model and by the FreeRate model had little inconsistency (RF=0.061 to 0.132, TT=0.0005 to 0.0034); the PMSF tree was more discrepant from the other two (RF=0.096 to 0.204, TT=0.0063 to 0.0121), yet this discrepancy was lower than revealed in other comparisons (Supplementary Fig. F.23c). This pattern was also indicated by hierarchical clustering (Supplementary Fig. F.23g, h). In the PCoA of RF distances, trees by the three models on the same dataset form compact clusters (Supplementary Fig. F.23e), whereas in the PCoA of tip distances, the PMSF trees had noticeable deviations from the Gamma and FreeRate trees (Supplementary Fig. F.23f). These observations suggest that the more complex and expensive PMSF method generated highly consistent topologies, but estimated slightly less consistent phylogenetic distances, comparing to the simpler models.

Collectively, this test also reveals that with our data set, the impact of taxon sampling on tree topology is notably larger than the impact of site sampling or model complexity, as evident in Supplementary Fig. F.23e across axis 2 (which explains 19.04% variance). For example, starting from the tree using 38k randomly selected sites with 1,000 taxa (small blue square), increasing site sampling to all 192k sites (small blue circle, a.k.a. "concat.al1k" in Fig. 6.3) resulted in RF=0.162, but increasing taxon sampling to all 10,575 taxa (big blue square, a.k.a. "concat.rand" in Fig. 6.3) resulted in RF=0.275 (also see Supplementary Fig. F.13).

Evaluation of species trees built using FastTree. While robust ML implementations like RAxML and IQ-TREE are computationally expensive and forced us to perform site downsampling, the faster alternative FastTree [261] allowed reconstruction of a CONCAT tree using all sites (192k in total). Since FastTree was used to reconstruct large-scale reference microbial phylogenies in several previous studies (e.g., [209, 246]), we compared the two methods in the context of our study. In particular, we compared species trees built using either FastTree or the robust method based on the conserved sites by a series of downsampled taxa.

Our results show that FastTree and the robust method produced similar topologies given the same input data, as long as the number of taxa is large (RF=0.111 to 0.408) (Supplementary Fig. F.12a). With different input data sets, both methods yielded relatively discrepant topologies (RF=0.438 ± 0.140 and 0.408 ± 0.139, respectively, mean and std. dev.) (Supplementary Fig. F.12c, d, upper left triangles), with FastTree trees being more discrepant (paired two-tailed *t*-test *p*=0.000247). In PCoA, input data dominantly determine the clustering pattern (PERMANOVA pseudo-*F*=7.117, *p*=0.001) of tree topologies, whereas method (RAxML vs. FastTree) has little effect (pseudo-*F*=0.679, *p*=0.752) (Supplementary Fig. F.12e). When considering the estimated phylogenetic distances among taxa, we observed a mixed effect. While input data set continued to impact the distribution of trees (pseudo-*F*=7.616, *p*= 0.001) (Supplementary Fig. F.12f), forming a clearly ascending gradient by the number of input genomes along axis 1 (which explains 62.19% variance), method also has a significant impact (pseudo-*F*=4.294, *p*= 0.025), clearly separating paired trees of each input data set on axis 2 (which explains 24.89% variance). The influences of input data and method on the tree distribution are comparable (RDA effect size: adjusted R^2=0.512 vs. 0.387, *p*= 0.006 and 0.004).

Therefore, despite the overall congruence in topology, there is a systematic bias between the two methods in estimating phylogenetic distances. Because our study has a strong focus on the evolutionary distances among microbial lineages, and considering that several previous studies associated FastTree with suboptimal likelihood scores [190, 381] and less accurate species tree [298], we decided to favor the robust method over FastTree when reporting our results. Conducting a comprehensive comparison between FastTree and RAxML / IQ-TREE is beyond the scope of this study. Nevertheless, we want to remind readers of this difference when interpreting the robust and FastTree trees, both of which were included in our data release.

F.1.3 Evaluation and curation of NCBI taxonomy

We evaluated the NCBI taxonomy [90] with reference to the ASTRAL tree. Of all 1,980 NCBI taxonomic terms with two or more representatives in our sampled genomes, only 1,219

(61.6%) terms are monophyletic. To further quantify the divergence between taxonomy and phylogeny, we computed the classification consistency [209] and the quartet score [297] of each term. The distribution of consistency scores reveals the imperfections of the taxonomy in reflecting the phylogenetically estimated relationships (Supplementary Fig. F.15a). Some large phyla were rejected consistently by different phylogenetic trees, pointing to potential inaccuracies in the taxonomy (Supplementary Fig. F.15c, see also Supplementary Note F.1.5). Using the automated taxonomy curation algorithm tax2tree [209], we reconstructed high-confidence taxonomic lineages for individual genomes and for internal nodes of the ASTRAL tree. This process does not create or modify taxonomic terms, but edits the assignments of genomes to existing taxonomic terms. When faced with strong signal of polyphyly for a taxonomic unit, tax2tree appends a numeric suffix to the taxonomic term for each clade (e.g., Fig. 6.1). This analysis established the taxonomy for 873 genomes that were unclassified at one or multiple taxonomic ranks by NCBI, and modified the existing taxonomy for 1,866 genomes (Supplementary Table F.3). Interestingly, at class, order and family levels, 19.36% of genomes defined as metagenome-derived received correction, while this ratio for genomes from isolates was much lower: 7.79% (one-tailed Fisher's exact test p-value=1.03e-23). This once more implicates the challenge in metagenome-assembled genome discovery and emphasizes the need for improved quality standards for this practice [34]. Source data are provided as a Source Data file. Annotations and curations are available from our data release.

F.1.4 Comparison with GTDB taxonomy and phylogeny

GTDB is a recent phylogenomics-curated taxonomy system for bacteria and archaea [246]. We compared our work with GTDB release 86.1. Among the 10,575 taxa in our phylogenetic analysis, 9,732 (92.0%) have matches in the GTDB taxonomy, and 8,042 (76.0%) of them are present in the GTDB phylogeny. We annotated our trees using the GTDB taxonomy (e.g., Supplementary Fig. F.16), and observed high overall congruence (Supplementary Fig. F.15b).

Among all 3,466 GTDB taxonomic units with two or more representatives, 3,403 (98.2%) are monophyletic in the ASTRAL tree. The congruence is also evident by directly comparing topologies of the GTDB phylogeny (composed of one archaea tree and one bacteria tree) and the ASTRAL tree (RF distance=0.185) (Fig. 6.3a, b). However, some differences in phylum-level organization and contents were observed (Figs. 3c, d, Supplementary Figs. F.3 and F.15d), and the ASTRAL tree appeared to have the fewest inconsistencies compared to the CONCAT trees using the global marker gene set (Supplementary Fig. F.15d). The differential inclusion of phylum-level classification units by the two works may contribute to this discrepancy. Further discussion of taxonomic units with reference to the GTDB trees and other published works is provided in Supplementary Note F.1.5. Source data are provided as a Source Data file. We included cross translations of genome identifiers and phylogenies of the two systems, and GTDB-based taxonomic curation of our genome pool in the data release.

F.1.5 Phylogenetic relationships of major taxonomic groups

We examined the placement of multiple important high-level (phylum and above) taxonomic groups in the species trees generated in this study. The ASTRAL tree (branch support: local posterior probability, or lpp) was used as the top-priority reference for the discussion, due to its stability and high resolution in deep phylogeny as discussed above and in the main text. The two CONCAT trees built using the robust ML implementation, based on either using conserved or random sites, were used for comparison in most discussions (branch support: rapid bootstrap, or xboot).

Archaea. The 669 representatives of the domain Archaea form a distinct clade in all three species trees (lpp=0.998 in the ASTRAL tree, xboot=100 in both CONCAT trees). The Archaea clade is split into the four currently accepted groups, namely Asgard, TACK, Euryarchaeota and DPANN [86, 50]. However, not all the groups are monophyletic, and this is particularly evident among the phylum Euryarchaeota (detailed below). Our trees do not support Asgard and TACK as sister groups (together as kingdom Proteoarchaeota, as proposed in [253]), despite the closeness of the two groups in the ASTRAL tree (detailed below).

Asgard. The recently discovered group of uncultivated archaea Asgard was considered to be close to eukaryotes and represent the archaea-to-eukaryote transition [377]. Our dataset includes eight representatives out of ten Asgard taxa from the original genome pool. Seven of them, representing the candidate phyla Lokiarchaeota (one taxon), Thorarchaeota (three taxa), and Heimdallarchaeota (three taxa), form a clade with moderate support (lpp=0.751, xboot=83 / 98) (a separation by "/" stands for conserved / random, same below) and reside in a relatively basal location in the Archaea lineage. In the CONCAT trees, this clade is sister to Marine Group II and III euryarchaeotes (13 and two taxa, respectively) (xboot=49 / 85), whereas in the ASTRAL tree, it is relatively independent (see below). In contrast to [377], our only representative of the candidate phylum Odinarchaeota is placed in a distant location, sister to a clade of four members of the candidate phylum Verstraetearchaeota (lpp=0.976, xboot=71 / 99), which is part of the TACK group. Therefore, tax2tree curation re-assigned Odinarchaeota to the TACK group. Meanwhile, two Asgard taxa were retained in the 1,000-taxon PMSF trees: one Thorarchaeota taxon is deeply nested within the TACK clade, with candidate phylum Bathyarchaeota (one taxon) being its sister (ufboot=99 / 100), whereas the other one, a Heimdallarchaeota taxon stands alone in a relatively basal position in the Archaea clade. We want to note the potential limitation in resolving Asgard placements due to its low availability of genome data.

TACK. The archaea TACK group (a.k.a., Proteoarchaeota) [121] was shown related to eukaryotes [253, 121] and placed as a sister group to Asgard in previous analyses [317, 50]. Members of the TACK group, including organisms under the phyla Crenarchaeota (169 taxa) and Thaumarchaeota (49 taxa), as well as the candidate phyla Bathyarchaeota (14 taxa), Korarchaeota (one taxon) and Verstraetearchaeota (four taxa), together with Odinarchaeota (see above), form a monophyletic clade with moderate support (lpp=0.88) in the ASTRAL tree. This topological pattern was also found in the CONCAT trees, but with weaker support (xboot=21 / 44). Further, in disagreement with ribosomal proteins-based results (e.g. [50]), all three trees in our study suggest that the TACK clade is sister to (lpp=0.979, xboot=55 / 92) the "Euryarchaeota_2 clade" (further discussed below). They together are sister to the Asgard group in the ASTRAL tree (lpp=0.917), although this proximity is not indicated by the CONCAT trees.

Euryarchaeota. The phylum Euryarchaeota includes most of the "conventional" archaea. This group (407 taxa) appears to be polyphyletic in all three trees, which is inconsistent with [50]. In the ASTRAL tree, this phylum splits into two clades: The major clade (Euryarchaeota_1) includes genomes of the class Thermoplasmata (25 taxa), Marine Group II (12 taxa), Methanomicrobia (132 taxa), Archaeoglobi (21 taxa), Halobacteria (99 taxa), Methanococci (19 taxa) and Methanobacteria (34 taxa). The minor group (Euryarchaeota_2) (lpp=0.752), comprising classes Thermococci (38 taxa) and Hadesarchaea (two taxa), plus the Arc I group archaea (eight taxa), forms a distinct sister cluster to the TACK group (see above). The CONCAT trees also show that Hadesarchaea and Thermococci are sister groups (xboot=21 / 45), and they together are sister to the TACK group (see above), but the Arc I group was placed in a different location, close to classes Methanococci and Methanobacteria. For comparison, the sister relationship between Thermococci and Arc I group was also supported in [246] and [50]. Arc I group is currently classified under the euryarchaeal class Methanomicrobia, but none of our trees support this hierarchical relationship. The position of the secondary Euryarchaeota clade is also supported by the PMSF trees on 1,000 taxa, which include three Thermococci and one Hadesarchaea taxa, forming a clade sister to the 19-taxon TACK clade (ufboot=100 / 99).

DPANN. The recently defined DPANN group of archaea [276] has five representatives in our analysis. In concordance with a recent study [50], our trees do not support the monophyly of this group. Two members of the candidate phylum Micrarchaeota form a distinct clade in all three trees. This group is basal to the entire Archaea clade in the ASTRAL tree (lpp=0.998). The candidate phyla Diapherotrites and Woesearchaeota each have one representative, and they form a clade with two unclassified archaea: GW2011_AR10 and GW2011_AR15. This clade is sister to the Micrarchaeota clade in the CONCAT trees with moderate support (xboot=67 / 63), but the two clades are not adjacent in the ASTRAL tree. In addition, the five representatives of the candidate order Altiarchaeales, which was recovered to be within the DPANN clade in previous studies [317, 50], form a clade nested within a big clade mainly composed of the orders Methanococcales and Methanobacteriales, and this clade is distant from the DPANN clades.

It should be noted that the taxon sampling of the DPANN group is sparse in this study compared to previous studies that focused on newly discovered organisms (e.g., [50]). This is mainly because the DPANN genomes have low numbers of detectable marker genes (67.81 ± 31.19, mean and std. dev.). As a consequence, only five out of 57 available genomes were selected using our genome subsampling protocol. (But see Supplementary Note F.1.7 for discussion of expanded DPANN sampling.) The proposed importance of DPANN in understanding the basal diversification of Archaea [358] calls for future improvements of our marker gene set.

CPR. The candidate phyla radiation (CPR) [40] comprises a large proportion of the bacterial diversity. Our trees include 1,454 CPR genomes, which form a single lineage with full support in all trees. Consistent with [50], the candidate phylum Wirthbacteria (one genome) is basal to the entire CPR clade, with full support in all trees. A clade comprised of the candidate phyla Peregrinibacteria [362] (60 taxa) and Abawacabacteria [11, 144] (one taxon) as sister groups (full support in all trees) was recovered as the second basal group in the CONCAT trees (full support) and as an early branching group, though not second basal, in the ASTRAL tree (full support). This pattern was not revealed in [50]. Most CPR taxa are grouped under two highly supported clades representing the superphyla Microgenomates [276] (a.k.a. OD1, 423 taxa) (lpp=0.913, xboot=97 / 96) and Parcubacteria [276] (a.k.a. OP11, 846 taxa) (lpp=1.0, xboot=99 / 100), respectively. The two clades are relatively derived and are not immediate sister groups. Thus the previous proposal of the superphylum Patescibacteria, comprised of Microgenomates, Parcubacteria, and the candidate phylum Gracilibacteria [276], is not supported [50]. Our sampling did not include any of the five genomes of Gracilibacteria, though, since they did not pass the quality filters. The candidate phylum Doudnabacteria [11] (19 taxa), was placed within the Parcubacteria clade in the ASTRAL tree and the random CONCAT tree, with weak support (lpp=0.621, xboot=60), a pattern consistent with previous work based on ribosomal proteins [11], but was basal to the entire Parcubacteria clade (xboot=100) in the conserved CONCAT tree. Overall, the relationships among major CPR candidate phyla were much more consistently resolved compared to phyla under non-CPR Bacteria (see below) (Supplementary Fig. F.3).

Non-CPR Bacteria (abbreviated as "***ncBacteria***" in this section). They form a monophyletic group in all trees based on global marker genes. This clade is highly supported in the ASTRAL tree (lpp=0.958) and in the random CONCAT tree (xboot=95) but less so in the conserved CONCAT tree (xboot as low as 29) (Fig. 6.3c). The CONCAT method struggled to resolve the relationships of the early branching ncbacterial clades, leaving poorly supported branches that were collapsed into polytomies in Supplementary Figs. F.5 and F.6. However, the ASTRAL tree provides remarkably higher resolution with moderate-to-high support of those basal relationships (Supplementary Figs. F.4 and F.9). In this tree, a clade is basal to the whole ncBacteria clade (full support), comprised of the phyla Thermotogae (35 taxa), Dictyoglomi (two taxa), and Caldiserica (two taxa), plus Firmicutes genera *Coprothermobacter* (three taxa) and *Thermodesulfobium* (one taxon). All of those taxonomic groups are featured by their thermophilic and anaerobic behavior. The basal placement of Thermotogae and other rooted groups within ncBacteria obviously support the hypothesis of an origin and early diversification of ncbacteria as (hyper)thermophilic anaerobes [369, 275].

Terrabacteria vs. "Hydrobacteria". Post the branching off of the (hyper)thermophilic bacteria clade in the ASTRAL tree, the ncbacteria clade split into two major clades (lpp=0.988). One (3,708 taxa) is mainly composed of taxa under the widely accepted term Terrabacteria, the largest group of ncbacteria that have shared adaptations to the terrestrial lifestyle [21]. Specifically, it contains the five originally suggested terribacterial phyla: Actinobacteria, Firmicutes (including

Tenericutes and Synergistetes), Cyanobacteria, Chloroflexi, and Deinococcus-Thermus [21], plus the more recently defined phylum Armatimonadetes (previously known as OP10) [330]. This clustering pattern was not revealed in [50] and [246]. The CONCAT trees inferred in this study also indicated mixed support/rejection for this clade (Fig. 6.3c, d). Multiple candidate phyla reside within the Terrabacteria clade, which help to further define their classification status. The other major clade (4,701 taxa), overlapping with the less commonly used term "Hydrobacteria" suggested by the same authors [21], contains the remaining ncbacterial diversity. The deep phylogeny of the Hydrobacteria clade reveals an interesting pattern of rapid diversification.

Aquificae vs. Thermotogae. The hyperthermophiles Aquificae and Thermotogae were conventionally determined as closely related groups (e.g., [50]) and together occupy the basal position of the ncbacteria clade [19, 21]. Our work, however, is consistent with that of [276] and found a clade containing the phylum Aquificae (17 taxa) and the candidate phylum Calescamantes [276] (a.k.a. EM19, seven taxa) (lpp=1.0, xboot=60 / 86), sister to a clade mainly comprised of class Epsilonproteobacteria (lpp=0.687, xboot=90 / 85) and distant from Thermotogae. Similar findings were obtained in some earlier comparative genome analyses of these groups [167, 117], while another study found no distinctive evolutionary relationship between the two groups [56], despite many members of them sharing similar ecology and physiology.

Synergistetes. The phylum Synergistetes (29 taxa, excluding one mis-classified taxon Synergistes sp. Zagget9) form a monophyletic clade in all three trees with full support and is proximate to several candidate phyla in the ASTRAL tree (lpp=0.787). However, in the CONCAT trees, the Synergistetes clade is paraphyletic to the thermophilic bacteria clade (see above) with low support (xboot=32 / 27). Previous studies suggested a close relationship between Synergistetes and Firmicutes, but had uncertainty in the placement of the Synergistetes clade relative to the latter [159]. Our trees suggest that Synergistetes is not an ingroup of Firmicutes, consistent with [225] but in contrast to [50].

Firmicutes/Tenericutes/Fusobacteria. The phylum Firmicutes has been widely reported to be a polyphyletic group, primarily because of the unstable positions of Tenericutes and/or Fusobacteria [361, 225, 144]. In our analysis, the 66 taxa of the phylum Tenericutes are nested within the Firmicutes clade in all three trees. However, this pattern is only credible in the ASTRAL tree (lpp=1.0), whereas in the CONCAT trees, the relevant branches have low support (xboot < 50). The Tenericutes taxa are para- or polyphyletic, mainly forming two clades, in close proximity to the Firmicutes class Erysipelotrichia (50 taxa). The taxa of the two groups cannot be clearly separated. It is remarkable that the Tenericutes clade has very long branch lengths compared to the remaining Firmicutes and the entire tree. These results show the non-determinacy of the hierarchical relationships between the two phyla. Unlike Tenericutes, the 36 taxa of the phylum Fusobacteria form a distinct cluster within the "Hydrobacteria" group in the ASTRAL tree (lpp=0.75), which is consistent with [225]. However, the CONCAT trees show that the Fusobacteria clade is nested within Firmicutes, sistering the Tenericutes-Erysipelotrichia clade, with low support (xboot=10 / 50). The instability of the class Clostridia, another Firmicutes group, has previously been noted [324, 120, 375], mainly as a result of misclassification of several species within the genus *Clostridium* [380]. In the ASTRAL tree, almost all the clades for class Clostridia (Supplementary Fig. F.4) have high support (lpp > 0.98), indicating that this tree can be an effective reference for resolving the problem of the classification of Clostridia.

Actinobacteria. Several orders in the phylum Actinobacteria, particularly Micrococcales and Pseudonocardiales, are widely known to be polyphyletic, and few efforts to rectify this problem using a combination of phylogenetic markers have been reported [236]. In our study, the phylum Actinobacteria was found as a monophyletic clade in the ASTRAL tree (lpp=0.986) and the random CONCAT tree (xboot=88). This finding is consistent with several previous studies [50] [246]. Recently, Parks, *et al.*, proposed to downgrade Nitriliruptoria to an order within the class Actinobacteria [246]. In our trees, however, the class Nitriliruptoria (one taxon) forms a distinct branch, well separated from the classes Actinobacteria and Acidimicrobiia.

Cyanobacteria/Melainabacteria. The candidate phylum Melainabacteria (17 taxa) is a recently discovered group of bacteria that are closely related to the phylum Cyanobacteria (a.k.a. Oxyphotobacteria, 295 taxa) but that lack the capability of photosynthesis [246]. Our trees support the members of Melainabacteria, plus 11 underclassified, metagenome-assembled genomes, as a fully supported monophyletic group, sister to the Cyanobacteria clade (lpp=1.0, xboot=100 / 98), which is also monophyletic (with full support). In contrast to [144, 50], our analysis did not recover it as a basal group to non-CPR Bacteria.

Chloroflexi. Members in the phylum Chloroflexi are model organisms for investigating a number of hypotheses related to the early evolution of photosynthetic life [308]. In all three trees, the 100 taxa of the phylum Chloroflexi form a single lineage (lpp=0.83, xboot=94 / 100). Our finding also suggests that the Chloroflexi group diverged during a similar period of time as the Cyanobacteria/Melainabacteria group (Supplementary Fig. F.25, see Supplementary Note F.1.6 for details), which is consistent with a recent study [308]. Furthermore, in this phylum, the order Chloroflexales is considered as the main phototrophic lineage that performed anoxygenic photosynthesis with a divergence time later than that of Cyanobacteria/Melainabacteria group. This observation does not support the hypothesis that anoxygenic photosynthesis preceded the development of oxygenic photosynthesis [306], in congruence with [308]. While the origin of photosynthetic life on the basis of the analysis of extant lineages is still unclear, the problem of undiscovered or extinct lineages further limits our understanding of evolution of phototrophy.

Spirochaetes. The basal position of the “Hydrobacteria” clade is occupied by four monophyletic lineages, represented by two cultured phyla – Fusobacteria (36 taxa) and Spirochaetes (135 taxa) – and two candidate phyla – Lindowbacteria (one taxon) and Aeriogibetes (three taxa). The evolutionary lineage of the phylum Spirochaetes in the ASTRAL tree and the random CONCAT tree is more consistent with [246], but contradictory to [50], which placed the phylum closer to the Proteobacteria. Further, in contrast to the view of Yarza, *et al.* [371], our trees do not support the classification of the phylum Spirochaetes into five lineages at the class level, but rather should be determined to have triphyletic subgroups (lpp=0.99): one containing the main order Spirochaetales (98 taxa), the second containing the order Brachyspirales (9 taxa), and the third containing the family Leptospiraceae of the order Leptospirales (27 taxa). The 43 taxa of the Spirochaetales family Borreliaceae form a shallow clade with a long stem, implicating a recent radiation.

PVC and FCB. The PVC and FCB superphyla groups form two monophyletic clades in all trees of life reported so far [21, 144, 246]. The topology of our trees also supports the divergence patterns reported earlier but provides a more robust position for an associated cluster of cultured and candidate phyla. Within this cluster, the phylum Gemmatimonadetes and the candidate phyla Glassbacteria, Eisenbacteria, Edwardsbacteria, Cloacimonates, Hyd24-12, and WOR-3 are closely related FCB (lpp=0.585), the candidate phyla Hydrogenedentes, Omnitrophica, Desantisbacteria, and Firestonsebacteria are closely related to PVC (lpp=0.99), and the rest, including the phylum Elusimicrobia and the candidate phyla Poribacteria and Coatesbacteria, form the root (lpp=1.0). While the robustness of our tree might be related to the number of selected marker proteins and/or the number of genomes used, the diversification of the different associated groups clearly suggests an evolutionary pattern for such divergence. For example, members of the phylum Gemmatimonadates can undergo both aerobic and anaerobic respiration, which enable them to adapt to an arid environment [329], while members of the phyla Chlorobi and Fibrobacteres are usually found under more strict anaerobic conditions [85].

Proteobacteria. The phylum Proteobacteria is the largest bacterial lineage of the rank, with 2,975 taxa in this study. The main subgroups of this phylum, particularly the classes Alphaproteobacteria, Betaproteobacteria, and Gammaproteobacteria are monophyletic, with the latter two sharing the same root. The class Epsilonproteobacteria (110 taxa) forms a sister clade (with full support) to a small clade comprised of deltaproteobacterial genera Desulfurella (one taxon) and Hippea (four taxa), then to the Aquificae-Calescamantes clade (see above). This pattern is consistent in all three trees, and is consistent with [246] but in disagreement with [144]. Our finding is also significant in the evolutionary point of view, as multiple Epsilonproteobacteria, particularly those isolated from deep-sea hydrothermal vents, meet their energy requirements through chemolithoautotrophy [45], a physiological condition related to the phylum Aquificae. The Epsilonproteobacteria-Aquificae clade is closely related to the class Deltaproteobacteria, which itself appears to be paraphyletic, with several other phyla such as Nitrospinae, Nitrospirae, and Thermodesulfobacteria nested within it. Parks, *et al.*, proposed to upgrade Epsilonproteobacteria and Deltaproteobacteria to a new phylum [246]. The distinctive placement of these two classes in our trees is roughly in concert with this proposal, though a more definitive study will be necessary.

F.1.6 Compatibility with geological timeline

We performed a series of divergence time estimation analyses to further demonstrate the efficacy of the 381 global marker genes in assessing the microbial evolutionary history. As revealed in Fig. 6.4, the evolutionary distance between Bacteria and Archaea was significantly shorter by using the global markers than by using the ribosomal proteins. Therefore, we focused on testing whether this observation is realistic, by projecting the species trees to the geologic timeline.

Maximum likelihood under a universal clock. Dating a phylogenetic tree of microbes has long been a challenge since few to no reliable fossil records are available to calibrate the tree [67, 128]. We performed a literature search and selected one calibration point that is among the most confident ones within bacteria and archaea:

Calibration 1: the origin of photosynthetic cyanobacteria. Specifically, it is the node that splits phylum Cyanobacteria and candidate phylum Melainabacteria, a recently discovered group of non-photosynthetic bacteria that are closely related to Cyanobacteria [71]. In our tree, the two sister clades have 295 and 28 taxa, respectively, with strong branch supports (further discussed in Supplementary Note F.1.5). It is widely accepted that the rise of oxygen in the Earth's atmosphere was a direct consequence of the evolution of photosynthetic bacteria, specifically, Cyanobacteria [315]. Recently, the Great Oxygenation Event (GOE) was precisely dated to 2.33 Ga (billion years ago) based on sulfur isotope signals [194]. In an independent study, the Cyanobacteria/Melainabacteria split was further estimated to be 2.5-2.6 Ga, using four calibrations based on well-accepted plant fossil records [307]. This range closely predates the GOE, indicating strong consistency with the aforementioned hypothesis of oxygenic photosynthesis evolution. Therefore, we adopted this range to constrain the Cyanobacteria/Melainabacteria split in the species trees.

We started with this single calibration, a simple assumption of one universal clock, and a maximum likelihood method which can be applied to the entire data set. The age of LUCA was estimated to be 4.1-4.2 Ga (in Hadean) by conserved sites, or 3.6-3.7 Ga (in Eoarchean) by random sites (Supplementary Table F.8). Either estimate is within the range consistent with the latest microfossil evidence [72] and in-silico estimations of life origination [203]. The split between CPR and non-CPR Bacteria took place 3.9 Ga (conserved) or 3.5-3.6 Ga (random). No later than 3.2 Ga (end Paleoarchean), all three major clades began to diverge (Fig. 6.6, Supplementary Fig. F.25). In contrast, using the ribosomal proteins, we obtained a very early estimate of the age of LUCA: 7 Ga (Supplementary Table F.8), which is inconsistent with the well-established age of the planet [62], whereas the divergence times of more derived lineages roughly agree with those by the global markers.

Impact of method, site sampling, site model and root placement. Comparative analyses suggest that the estimated ages were mainly influenced by gene and site sampling, whereas the impact of the tree-building method was minimal (Supplementary Table F.8). Considering the potential impact of root placement on the analysis, we moved the root from the midpoint of the Archaea-Bacteria branch to the first and third quarters, and obtained consistent results (Supplementary Table F.8). We then examined the impact of site model (PMSF vs. Gamma) on the 1,000-taxon trees (Supplementary Table F.8). For global markers, the difference is minimal. The age of LUCA estimated by random sites agree with the full tree (3.7 Ga), while that by conserved sites is slightly earlier (4.5 Ga), likely an impact of taxon downsampling (discussed above). For ribosomal proteins, the age of LUCA was further pushed to 9.2 Ga by PMSF from 7.5 Ga by Gamma.

Alternative calibrations. We tested the compatibility of multiple other calibration points and ranges with the photosynthetic cyanobacteria-based estimation, although these hypotheses are usually controversial or less precise (with lower bound only).

Calibrations 2 and 3: The origin of photosynthetic eukaryotes. The widely adopted endosymbiotic theory [382] suggests that eukaryotic organelles originated from symbiotic prokaryotes. The earliest fossil of photosynthetic eukaryote with relatively evident morphological characteristics, *Bangiomorpha pubescens* (a red alga), was recently precisely dated to 1,047 +13/-17 Ma [109]. Therefore we used the age 1.03 Ga to define the lower bounds of postulated bacterial and archaeal lineages from which organelles evolved through endosymbiosis. Specifically, it is commonly agreed that plastids evolved from cyanobacteria [223], although the specific cyanobacterial lineage is under debate (e.g., [238, 259]). Therefore we placed this calibration at crown Cyanobacteria.

On the other hand, it has been long suggested that mitochondria evolved from an alphaproteobacterial lineage, most likely Rickettsiales [280]. However, a recently study placed the mitochondrial origin at a proteobacterial lineage that branched off before the diversification of alphaproteobacteria [204]. We tested both theories, by placing the calibration at either crown Alphaproteobacteria (which has 893 taxa) or the split between Alphaproteobacteria and other proteobacteria (mostly beta- and gammaproteobacteria).

Calibration 4: The origin of akinetes-forming cyanobacteria. Several groups of extant cyanobacteria under families Nostocaceae and Stigonemataceae (both belong to order Nostocales) have the capability of forming environmental stress-resistant cells: akinetes [341]. Fossil akinetes (referred to as *Archaeoellipsoides*) have been recorded from a wide time period, most frequently between 1.4 Ga and 1.65 Ga [341]. The relationship between those records and modern Nostocales species remains controversial [43]. Despite being a frequently used calibration (e.g, [65]), some authors chose not to adopt it considering the controversy (e.g., [25]), and some found it to strongly impact age estimation (e.g., [198]). In our tree, order Nostocales (54 taxa) is monophyletic and nested within the Oscillatoriales clade, which is roughly consistent with [341]. We sequentially constrained the origin of the Nostocales clade with four representative ages of fossil akinetes: 1.2 Ga [138], 1.5 Ga [114], 1.9 Ga [113] and 2.1 Ga [10].

Calibration 5: The origin of aphid-*Buchnera* symbiosis. *Buchnera aphidicola* is the primary obligate symbiont of aphids (Aphidoidea) [252]. This close relationship was estimated to originate from 84-164 Ma [73], as evident by the radiation of fossil aphids and the implication from a geological thermal shift. This estimate is roughly consistent with more recent studies on larger scopes (e.g., [154]). Some authors (e.g., [65]) applied this calibration to the split between *Buchnera* and *Wigglesworthia* (obligate symbionts of a different host: tsetse fly). In our robust taxon sampling, a *Candidatus* Tachikawaea gelatinosa [160] taxon is slightly more closely related than *Wigglesworthia* to the eight-taxon *Buchnera* clade, however considering that it has not been rigorously studied, we still placed the calibration at the *Buchnera/Wigglesworthia* split, and we used either 84 Ma or 164 Ma to define the lower bound of it.

Our results (Supplementary Table F.9) show that the estimated ages of LUCA and non-CPR Bacteria remained largely consistent when either or both the photosynthetic eukaryotes calibrations and the aphid-*Buchnera* symbiosis calibration, with all their variants, were included in addition to the photosynthetic cyanobacteria calibration. However when the akinetes-forming cyanobacteria calibration (with any of the four variants) was introduced, it strongly pushed the estimations backward to an unlikely range. These results provide new information for paleobiological discussions.

Bayesian inference with alternative models. To validate and further strengthen the findings from maximum likelihood and the simple assumption of one clock, we analyzed the data using the more robust Bayesian inference method, with alternative clock models (strict or relaxed). The computational challenge forced us to downsample data to 5,000 sites by 100 taxa (the impact of downsampling was discussed in Supplementary Note F.1.2), the latter of which was selected to maximize the representation of deep phylogeny, but also to include sufficient sampling around the calibration point. Specifically, seven Cyanobacteria and three Melainabacteria taxa were included.

We tested two alternative prior distributions of time constraints. First ("narrow"), we adopted the estimated 2.5-2.6 Ga range (see above), and specified a normal distribution with mean=2.55 and std. dev.=0.025, so that 95% probability falls with this range. Next, we explored paleogeological evidence and alternative theories of cyanobacteria evolution, and specified a more relaxed constraint ("wide"):

Calibration 1 rev. Robust isotopic records have been found indicative of free oxygen in ocean or atmosphere around 3.0 Ga [60, 258, 234], while the earliest putative evidence was dated to 3.23 Ga [296]. The connection between early signs of oxygen with photosynthetic cyanobacteria has long been suggested [195], although the relationships among early oxygen, phototrophy, filamentous microfossils and ancestral cyanobacteria remain much debated, and usually questioned by recent studies [300, 49, 242, 221]. Here we adopt a treatment analogous to Shih *et al.* [307], by placing a soft upper bound at 3.0 Ga.

Accordingly, we specified a lognormal distribution, with offset=2.33, which is the date of GOE (see above), mean=0.22, so that mean + offset=2.55, which is in the midpoint of the estimated range (see above), and std. dev.=0.268, so that 95% probability falls before 3.0 Ga, when free oxygen was evident (see above) (Plus, 97.5% probability falls before 3.23 Ga, see above).

Our results (Supplementary Table F.10, Supplementary Fig. F.26) show that the estimated ages of LUCA were close between alternative clock models (strict vs. relaxed) and time constraints (narrow vs. wide), and supported the results based on on full-scale trees. We also calculated the coefficient of variation (C.V.) of clock rate under the relaxed clock model, a measurement of how “clock-like” the data are [76]. The C.V. by using the global markers (despite randomly downsampled to 5,000 sites) was ∼0.175, showing a modest deviation from a universal clock. Meanwhile, the C.V. by using the 30 ribosomal proteins was ∼0.254, suggesting a larger violation.

Taken together, we demonstrated that the microbial evolution dated using the 381 global marker genes and our species tree correspond well with the current paleobiological and geological evidence and theories. In contrast, the ribosomal proteins, which tend to overestimate the evolutionary distance between Bacteria and Archaea (see main text), consistently resulted in LUCA age estimates far older than Earth formation. This implicates a strongly accelerated evolution in the ribosomal proteins during the Bacteria-Archaea split. Therefore, we suggest that future researchers take caution when attempting domain-level divergence time estimations using a handful of “core” genes such as the ribosomal proteins. Although more comprehensive studies will be required, our analysis has indicated value of using the global marker genes for more accurate divergence time analysis. Nevertheless, we do not recommend treating our result (Supplementary Figs. F.25 and F.26) as a precise time table for microbial evolution, considering the simplicity of method and the sparsity of reliable and accurate calibrations.

F.1.7 Phylogenetic analysis with latest genome availability

We collected bacterial and archaeal genomes from NCBI RefSeq and GenBank on May 23, 2019. From this updated genome pool, we examined phylum-level classification units as defined by the latest NCBI taxonomy (released on June 1, 2019, which is after RefSeq 94) and GTDB taxonomy (version 4, released on June 19, 2019, indexed to RefSeq 89). For phyla that are absent, or represented by less than three genomes in the current set of 10,575 genomes, we selected new genomes with highest number of marker genes (must be no less than 100) to make the sampling up to three within each phylum. Genomes with CheckM contamination score larger than or equal to 5% were excluded. This process added 187 new genomes, representing an added or updated set of 52 NCBI phyla and 66 GTDB phyla.

We performed phylogenetic reconstruction with the 187 genomes added to the dataset, totaling 10,762 genomes. The procedures are largely consistent with the ASTRAL and CONCAT methods as described above, with several modifications to reduce computational expense (see Methods). Importantly, the same set of 381 marker genes and the same set of up to 100 most conserved or randomly selected sites per gene were used, granting comparability with the main analysis.

The resulting phylogenetic trees are highly consistent with the main results. In the ASTRAL tree we observed the highest consistency with the main ASTRAL tree (RF=0.035) (Supplementary Fig. F.27), while the two CONCAT trees using either most conserved or randomly selected sites also show high consistency with the corresponding CONCAT trees in the main analysis (RF=0.122 and 0.099, respectively). All three trees support the separation of Archaea, CPR and non-CPR Bacteria. The domain-level evolutionary distances are also highly close to the main results (Supplementary Table F.11). Therefore, our main findings hold with the up-to-date genome data.

The newly added genomes provide several insights. First, in the ASTRAL tree a new clade is placed at the base of the non-CPR Bacteria clade, consisted of three genomes classified as phylum UBP7 in GTDB. This placement is consistent with Parks *et al.* [246] in that it is the most CPR-proximal clade. However the CONCAT trees lack resolution at the base of the non-CPR Bacteria clade to reveal this relationship (see also Supplementary Figs. F.3 and F.8). Second, the previously underrepresented DPANN group (five taxa) was expanded, and revealed the same phylogenetic pattern (see Supplementary Note F.1.5). Specifically, the main clade residing at the base of the Archaea clade now contains six DPANN genomes and two unclassified genomes, and the secondary Micrarchaeota clade now has four taxa and is still separated from the main clade.

F.2 Supplementary figures

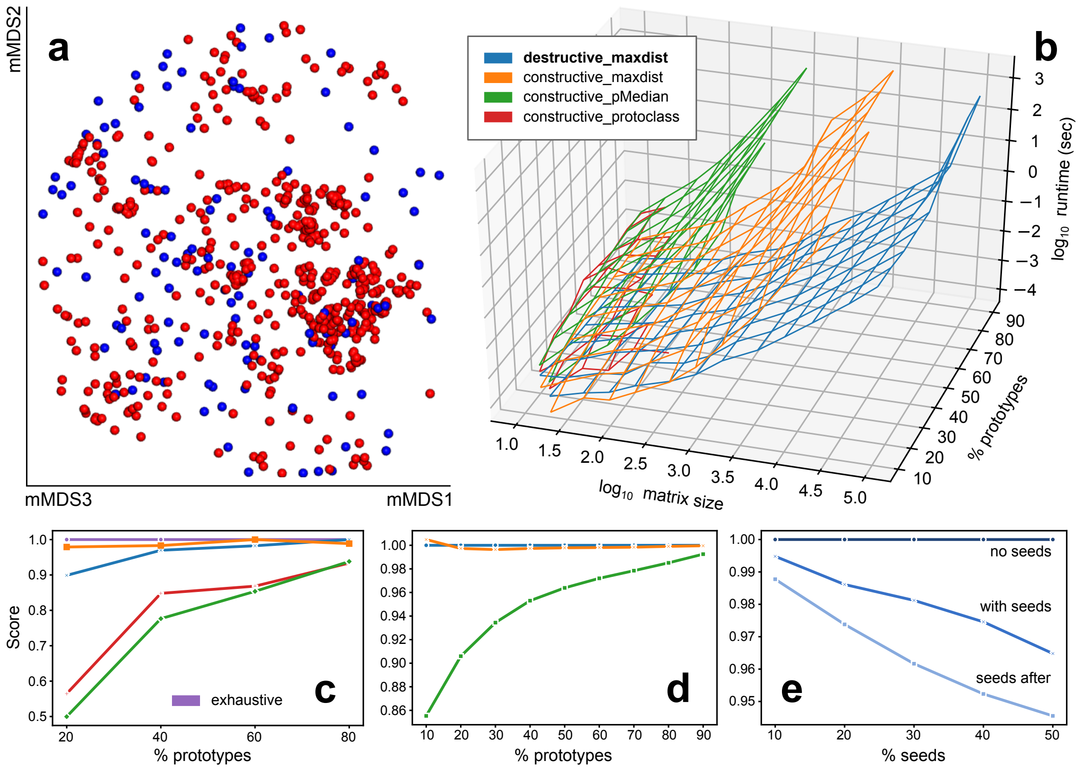


Figure F.1: Prototype selection for maximizing biodiversity included by fixed number of genomes. **a**. Visual effect of the final result of the genome subsampling workflow: metric MDS plot of the genome distance matrix, showing selected genomes (blue) vs. remaining ones (red). Despite that the distribution of genomes is highly uneven, this statistical approach delivered an evenly-distributed subset of genomes. Considering computational challenge and visualization purpose, this plot shows 1,000 genomes randomly sampled from all 86,200 genomes, of which, 112 belong to the 10,575 genomes selected for phylogenetic reconstruction. **b**. Runtime comparison of four alternative heuristics to solve the prototype selection problem (detailed in Supplementary Note F.1.1), of which destructive_maxdist was eventually used to subsample genomes in this work. The x-axis is the size of the randomly generated distance matrix: $n = |D|$, the y-axis is the amount of prototypes to select: k, given in ratios of n, and the z-axis is runtime in seconds. Execution time was limited to one hour at most. The runtimes for constructive_protoclass were trimmed off early because it could not find solutions for the given k with large datasets. **c**. Score (sum of pairwise distances among selected data points) normalized by that of the exact best solution (as computed using exhaustive search) vs. ratio of prototypes, on a small distance matrix with n=25. **d**. Score normalized by that of destructive_maxdist vs. ratio of prototypes, on a moderate-size distance matrix with n=1000. **e**. Scores of destructive_maxdist at n=1000 and k=20%, when randomly selected seeds (r, given in ratios of k) were provided ("with seeds"), as normalized to that when no seeds were specified ("no seeds"). The third curve, "seeds after" was computed when the same set of seeds were removed from the distance matrix prior to prototype selection, and then added back to the selection after the operation. In another word, it bypassed the "seeds" function implemented in the destructive_maxdist algorithm. Source data are provided as a Source Data file.

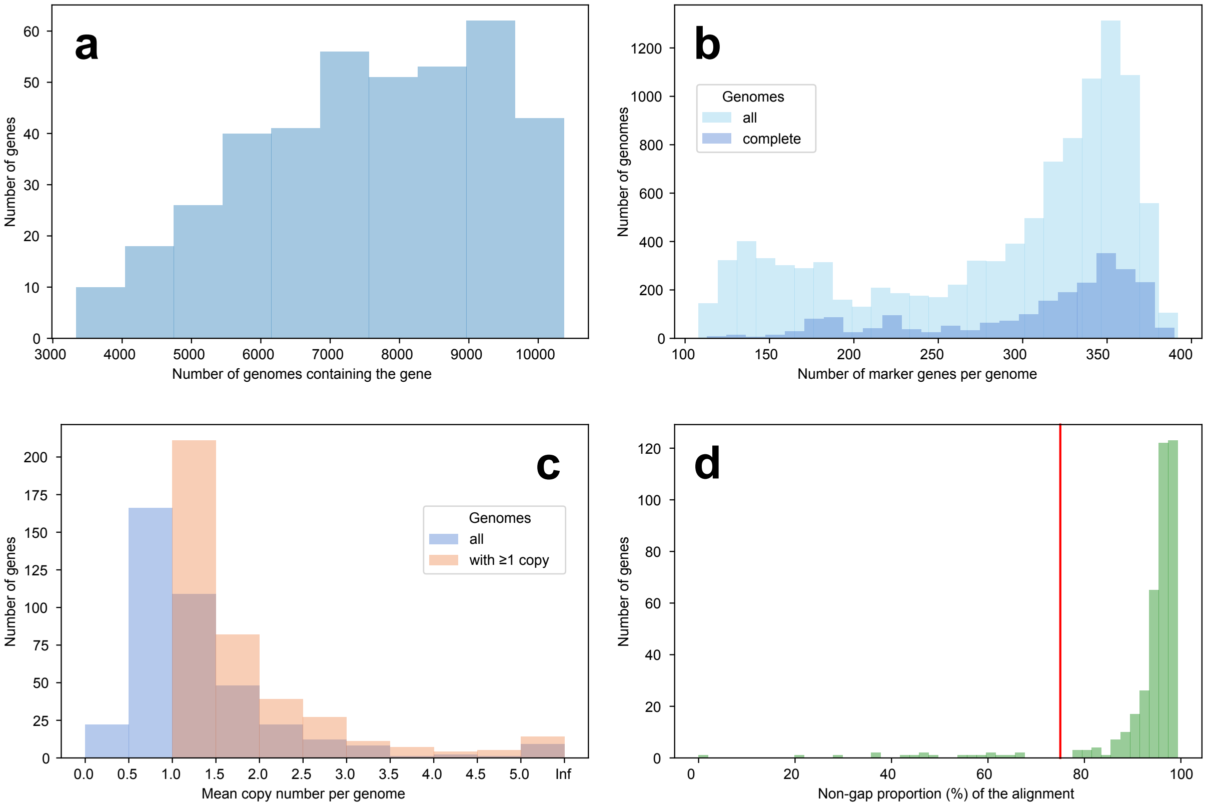


Figure F.2: Statistics the 400 marker genes in the 10,575 sampled genomes. **a**. Distribution of the number of genomes where individual genes were identified. **b**. Distribution of the number of identified marker genes per genome. “Complete” is a subset of all genomes, which were marked as “Complete Genome” or “Chromosome” by NCBI. **c**. Distribution of mean copy number per genome of each marker gene. The “copy number” is the count of USEARCH hits at an E-value threshold of 1e-40 during the PhyloPhlAn marker gene discovery. **d**. Distribution of the proportion of non-gap sites in the multiple sequence alignments of individual marker genes. The red vertical line indicates the threshold we chose based on observing this distribution pattern. Nineteen marker genes below this threshold were dropped, leaving 381 for the subsequent phylogenetic analysis. Source data are provided as a Source Data file.

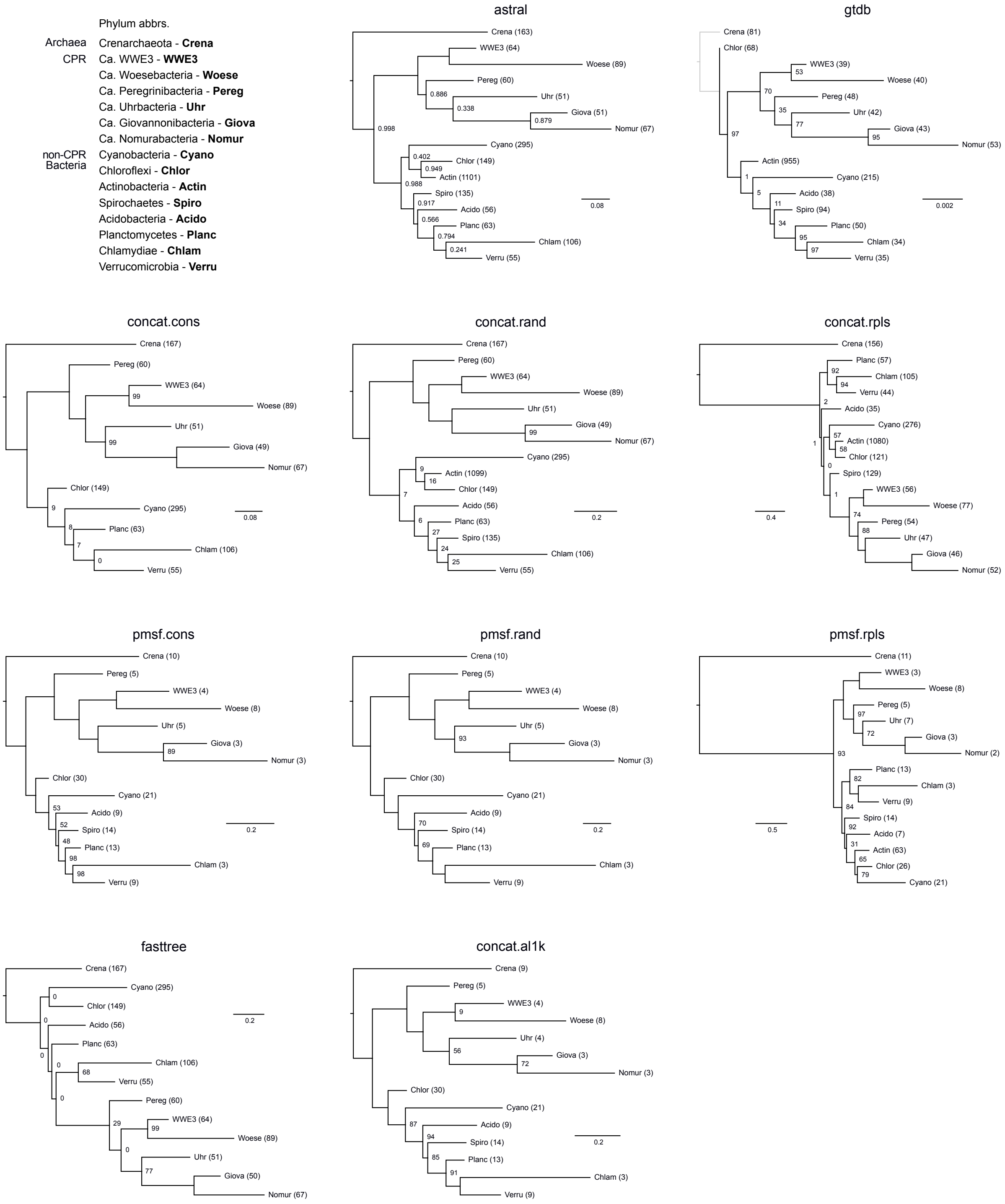


Figure F.3: Phylum-level relationships revealed by multiple species trees. Nine species trees reconstructed in this work plus the previously published GTDB release 86.1 tree are displayed (see Fig. 6.3). The phyla were selected from the tax2tree-curated NCBI phyla based on the ASTRAL tree. Fifteen most specious phyla which had no significant violation of monophyly according to tax2tree were selected. For each of the other nine trees, the same 15 phyla were

selected, but any of them was omitted if it violated monophyly based on the tree-specific tax2tree curation. Only the LCA of each phylum is shown, while all descending branches were pruned. Numbers in parentheses represent the number of descendants under each clade. Node labels represent branch support values (see Fig. 6.3). Nodes without labels were fully supported. The branch length scales are in the unit of number of substitutions per site. For display purpose, the branch lengths of the ASTRAL tree were estimated using conserved sites (same as in Fig. 6.1). Also for display purpose, the GTDB Archaea tree and Bacteria tree were artificially connected by a grey line which bears no information of topology or branch length. Source data are provided as a Source Data file

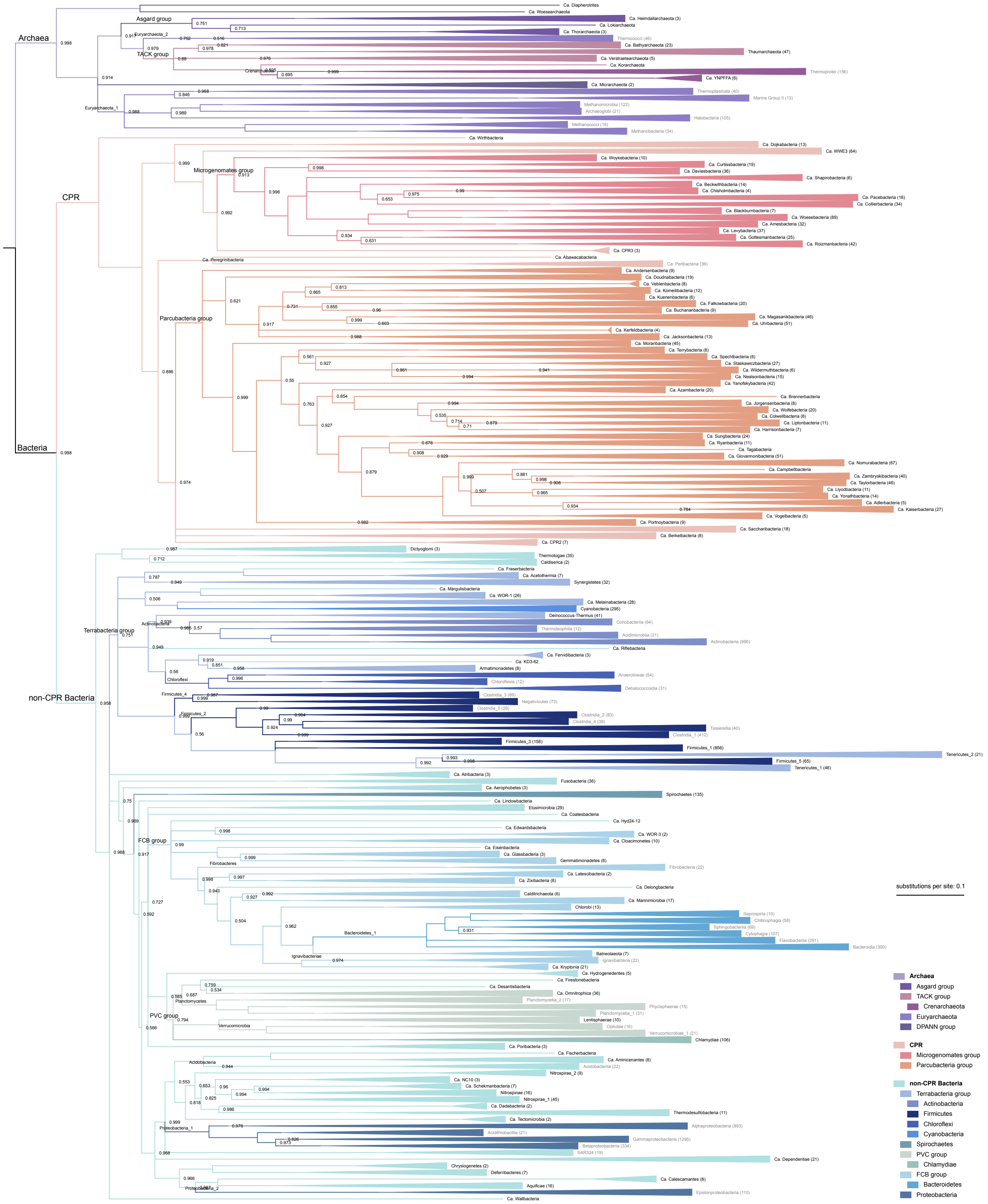


Figure F.4: The ASTRAL summary tree rendered in rectangular layout, collapsed to class level. The displayed features are consistent with Fig. 6.1. The triangles represent collapsed clades, with length equal to the longest branch in the clade. Node labels represent local posterior probability (lpp) of the corresponding branch. Labels are omitted at fully-supported (lpp = 1.0) branches. Source data are provided as a Source Data file.

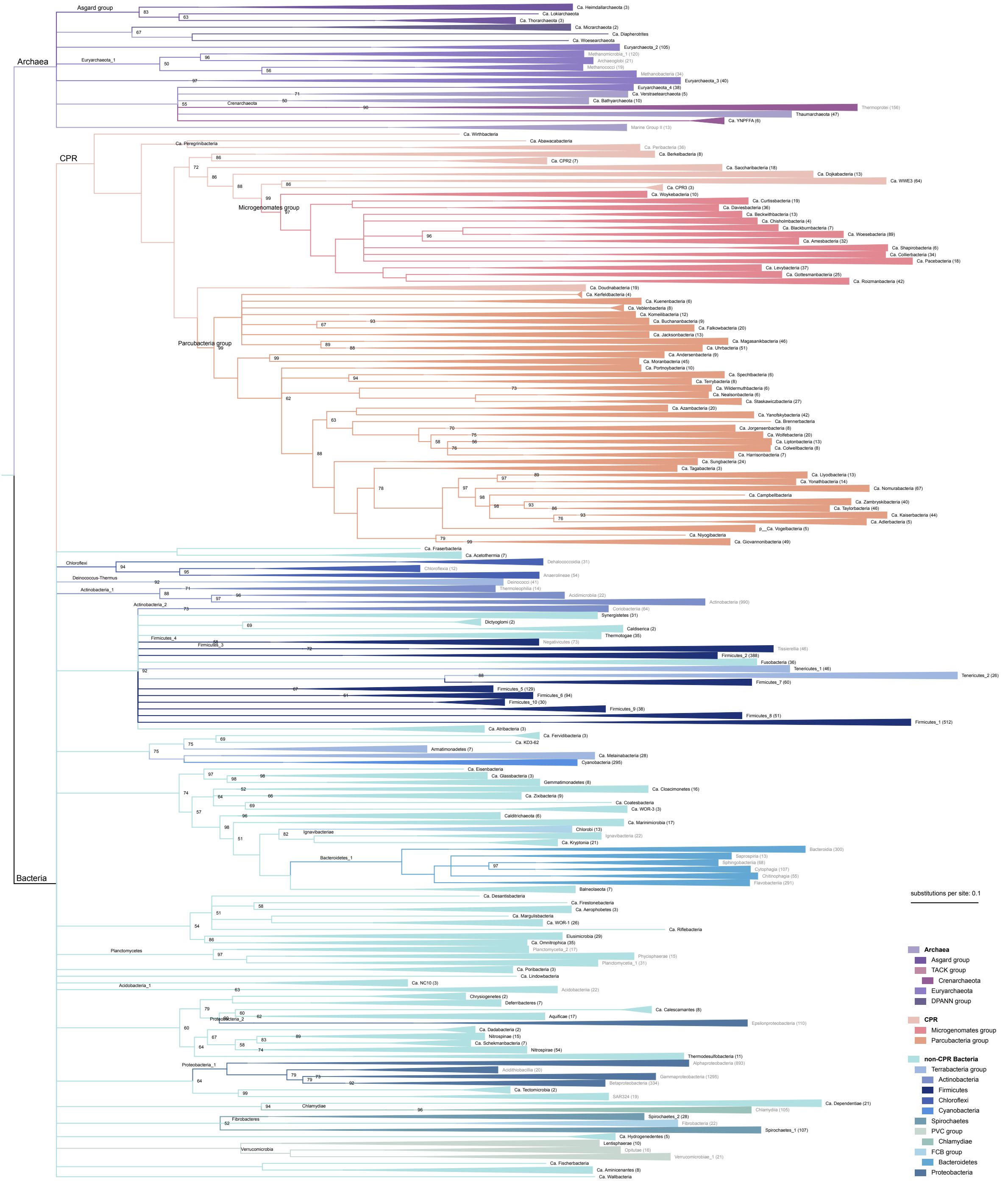


Figure F.5: The RAxML concatenation tree based on the 100 most conserved sites per gene, rendered in rectangular layout, collapsed to class level. Node labels represent rapid bootstrap support values (out of 100). Labels are omitted at fully-supported branches. See Supplementary Fig. F.4's caption. Source data are provided as a Source Data file.

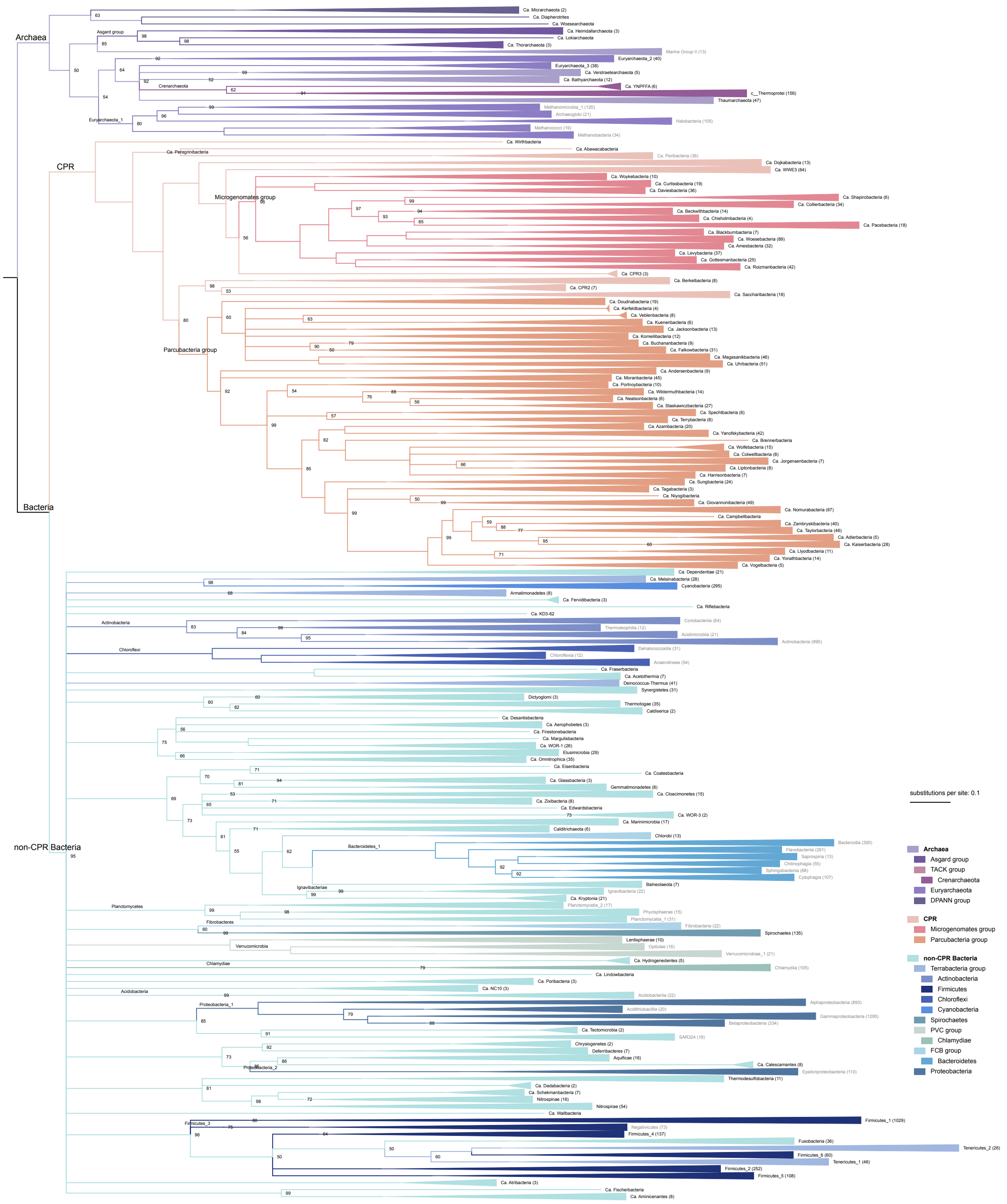


Figure F.6: The RAxML concatenation tree based on the 100 randomly selected sites per gene, rendered in rectangular layout, collapsed to class level. Node labels represent rapid bootstrap support values (out of 100). Labels are omitted at fully-supported branches. See Supplementary Fig. F.4's caption. Source data are provided as a Source Data file.

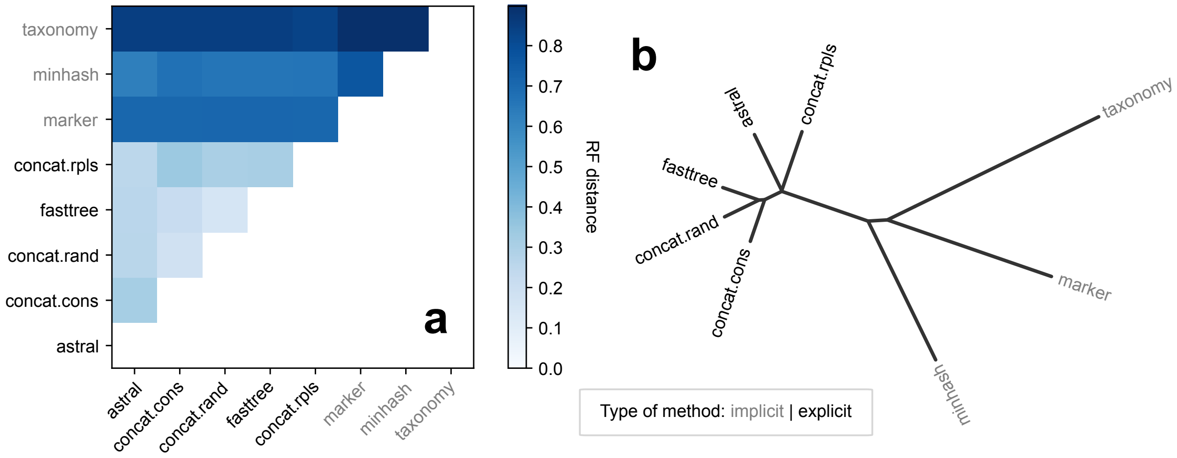


Figure F.7: Comparison of topologies of species trees built using explicit and implicit methods. **a**. Heatmap of RF distance matrix. **b**. Hierarchical clustering of RF distance matrix. “taxonomy”: NCBI taxonomy hierarchy; “minhash”: neighbor-joining (NJ) tree based on the Jaccard distance matrix calculated using the MinHash signature of genomes; “marker”: NJ tree based on the Jaccard distance matrix calculated using the presence / absence of the 400 marker genes in genomes; “concat”: phylogenetic trees built using the conventional gene alignment concatenation strategy; “astral”: phylogenetic tree built using the gene tree summary method ASTRAL; “cons”: 100 most conserved amino acid sites per each of the 381 marker genes; “rand”: 100 randomly selected sites per gene; “fasttree”: all sites, but tree was inferred using FastTree (the other concat trees were inferred using RAxML); “rpls”: 30 ribosomal proteins instead of the 381 marker genes. Source data are provided as a Source Data file.

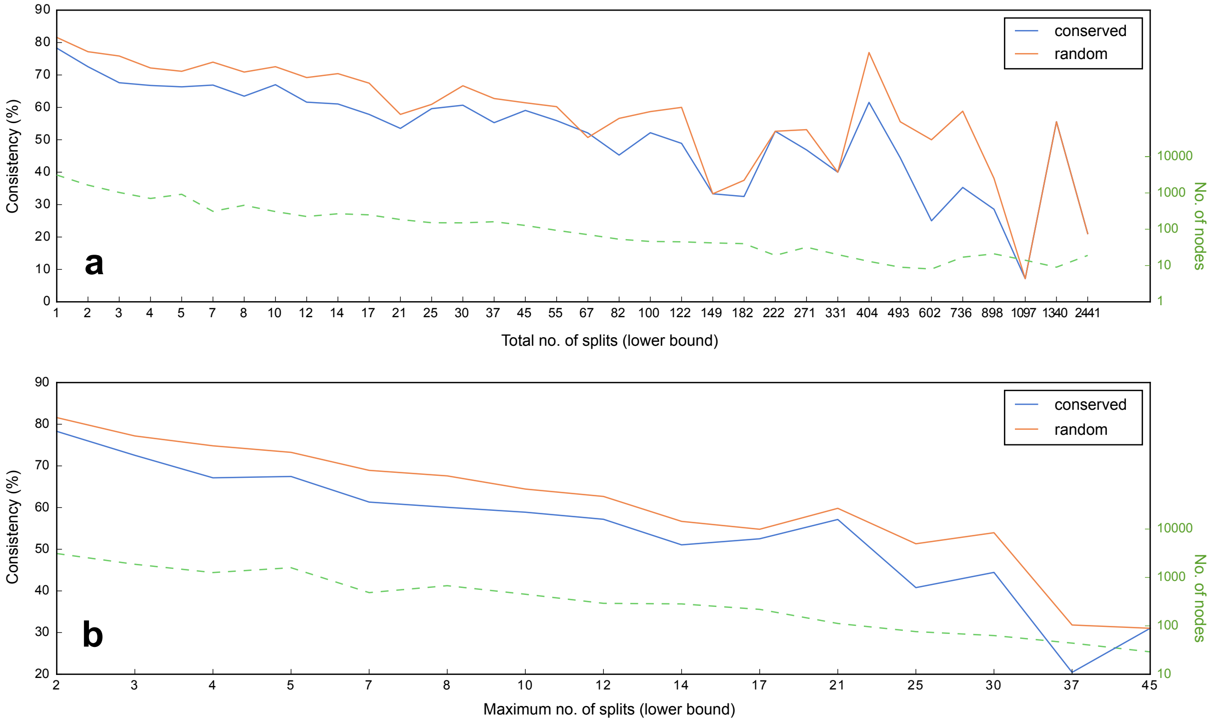


Figure F.8: The consistency score (*y*-axis) is the proportion of internal nodes in tree 1 that can be matched to a node in tree 2 which has exactly the same set of descendants. We measured the phylogenetic depth (*x*-axis) using two metrics: **a**. the total number of splits in the clade. This metric was introduced in 96 as the “split depth”. The *x*-axis was binned on a roughly logarithmic scale, as determined by Python code: sorted(set(int(math.exp(x/5)) for x in list(range(40)))). Bins with population size (number of nodes) less than five were merged into the next bin. **b**. the maximum number of splits from any tip to the node. The *x*-axis was binned by Python code: sorted(set(int(math.exp(x/5)) for x in list(range(20)))). The per-bin population sizes are indicated by the red dashed lines. Source data are provided as a Source Data file.

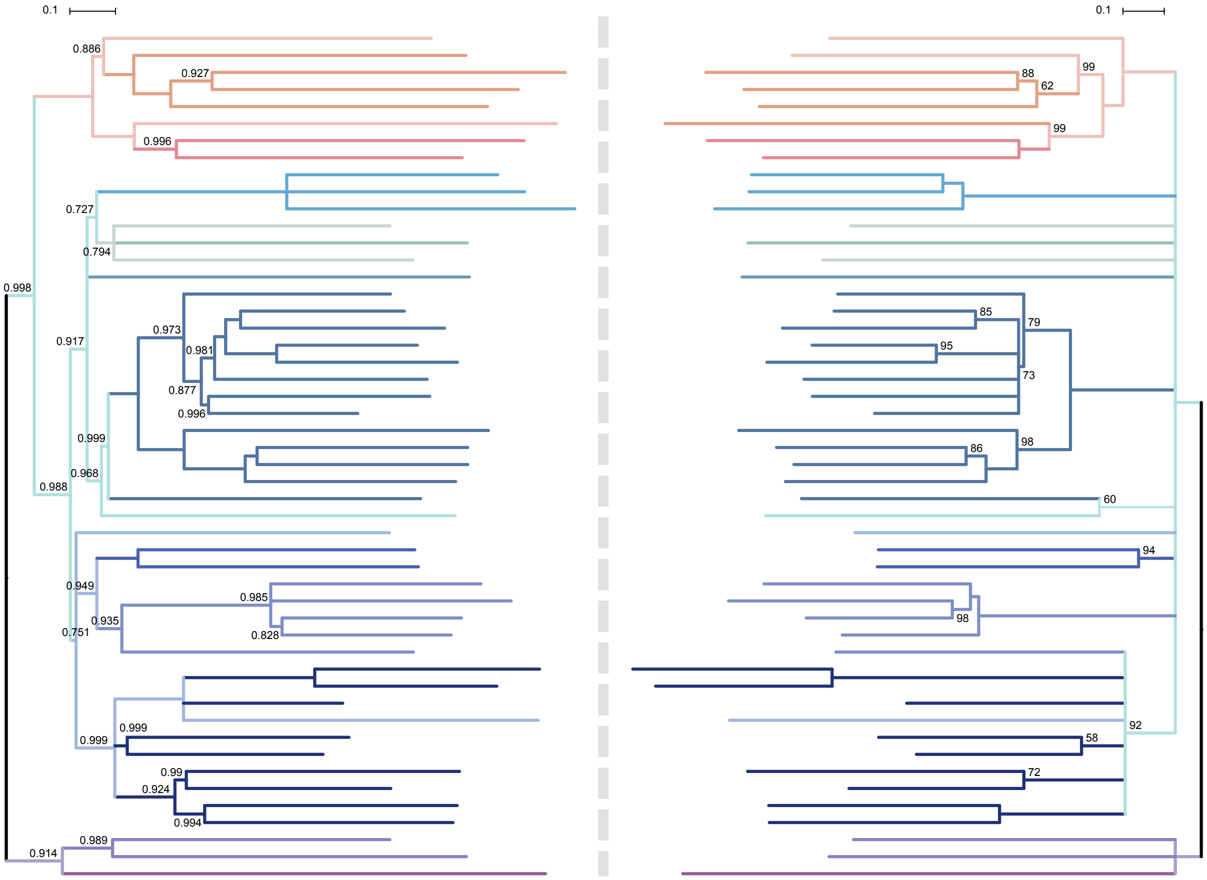


Figure F.9: Back-to-back comparison between the ASTRAL tree (left) and the CONCAT tree (right). Both used the conserved site sampling. Low-support branches were collapsed from the two trees to retain the same number of internal nodes per tree. The two trees were then collapsed to 50 shared clades with 50 or more descendants each. A tanglegram was generated to align the clades. Non-full branch support values (local posterior probability for ASTRALand rapid bootstrap for CONCAT) were annotated as node labels. The branches were colored using the same color scheme as in Fig. 6.1. Source data are provided as a Source Data file.

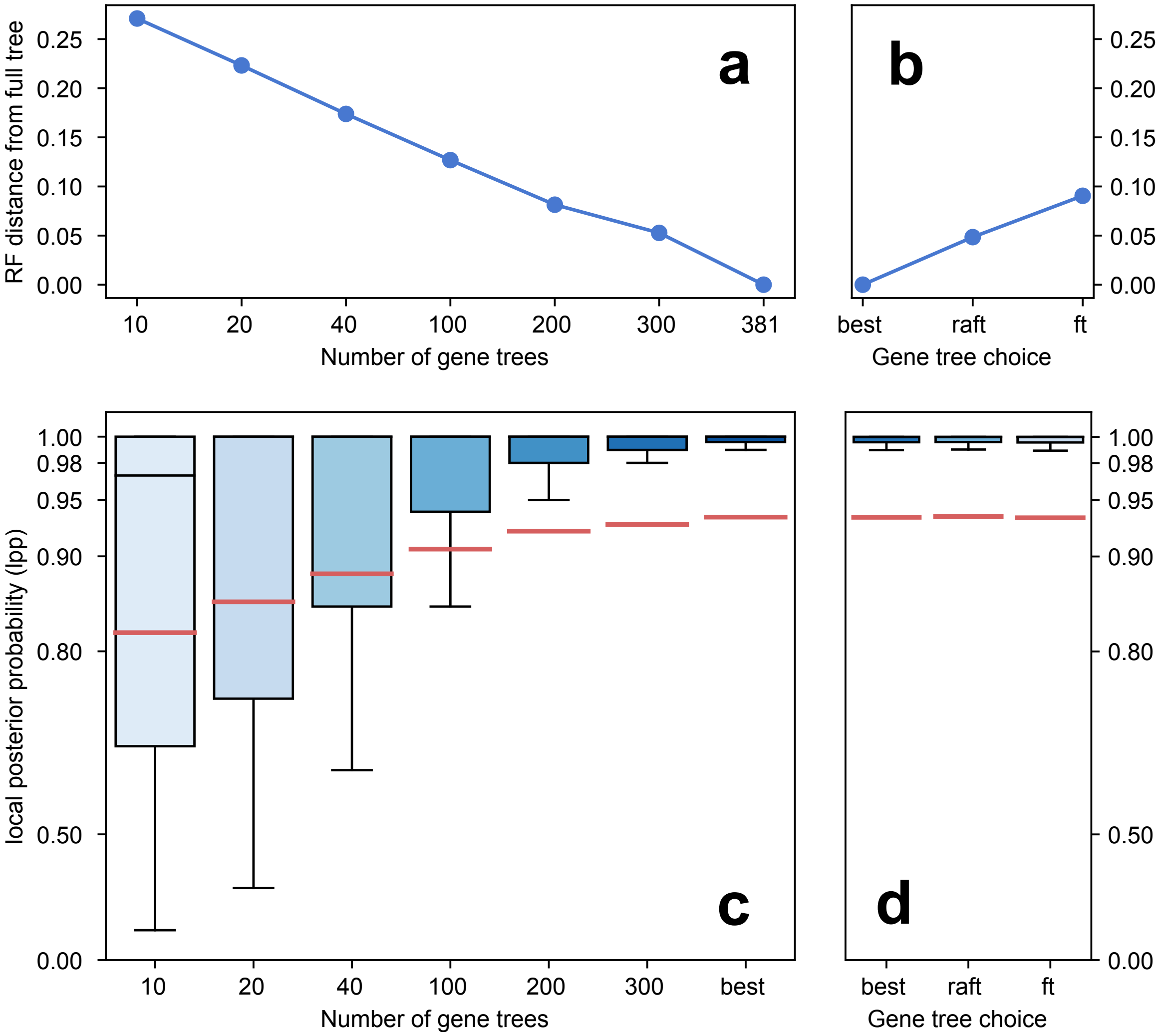


Figure F.10: Comparison of ASTRAL species trees built from differential quantity and quality of gene trees. **a, c**. Series of numbers of gene trees randomly sampled from all 381 gene trees. **a,d**. All gene trees, built and selected using different methods: "ft": gene trees inferred using FastTree; "raft": gene trees inferred using RAxML, with the FastTree trees as the starting trees; "best": for each marker gene, select one tree which has the highest likelihood score from three RAxML runs: one by the FastTree starting tree and other two by random seeds. **a, b**. RF distance from the full-scale reference tree (i.e., "381" in **a** or "best" in **b**). **c, d**. Distribution of branch support values (local posterior probabilities, or lpps). The red lines represent means. The y-axis is in exponential scale. Source data are provided as a Source Data file.

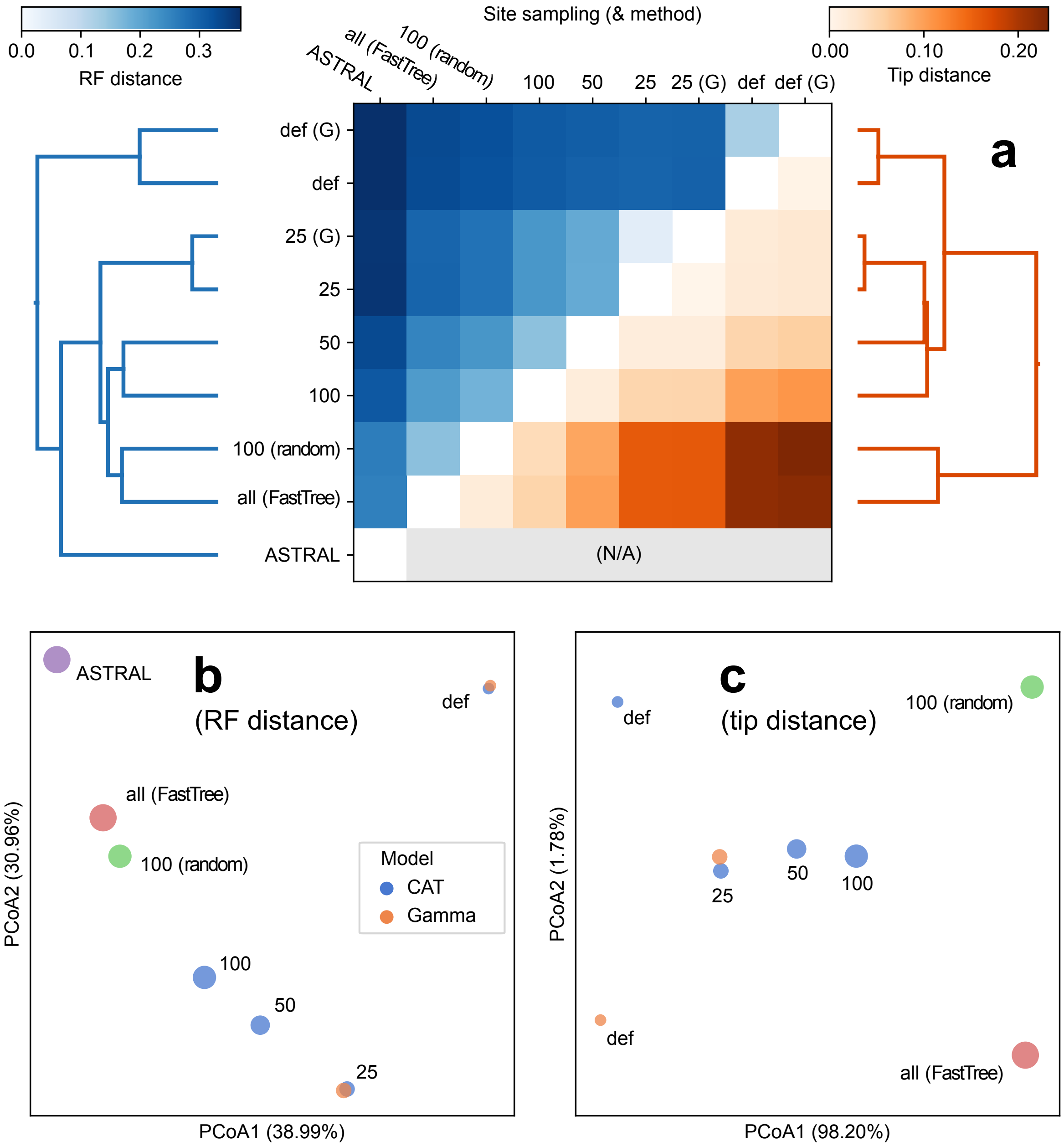


Figure F.11: Comparison of CONCAT trees built using different site sampling strategies. **a**. Heatmap and hierarchical clustering of RF (blue) and tip (orange) distance matrices. The tip distance measures the discorrelation between the two phylogenetic distance matrices among taxa in two trees (see Methods). The full-length marker gene alignments were subsampled based on maximum conservation, at a series of: PhyloPhlAn default ("def"), which approximately yielded 12 sites per gene, and 4.5k sites in total; then 25 sites per gene (9.5k in total), 50 sites per gene (19k in total), and 100 sites per gene (38k sites in total). For def and 25, we were able to perform RAxML tree search under the Gamma model, so the resulting trees were included in this comparison, but for 50 and 100, the use of Gamma model was prohibited by computational

challenge. For comparison, we included a tree built on alignments randomly subsampled to 100 sites per gene ("random"), and a tree built on all sites without subsampling, but using FastTree ("all"). Finally, we included the ASTRAL tree, based on gene trees built using all sites, as a reference for comparing topology, but it was not included in the comparison of distances, as the branch lengths inferred by ASTRAL are not comparable to those by CONCAT. **b** and **c**. PCoAs of RF and tip distance matrices, respectively. Source data are provided as a Source Data file.

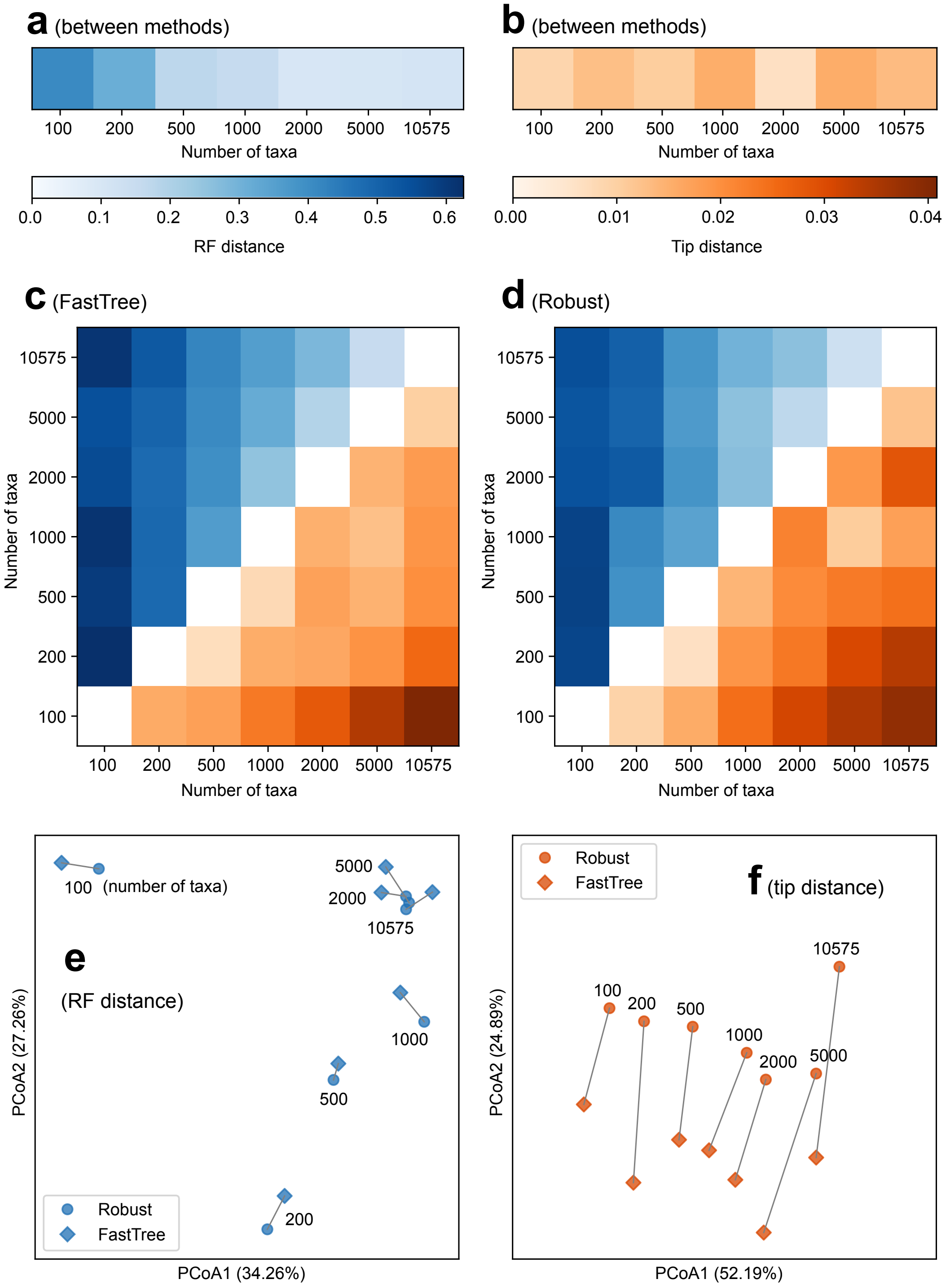


Figure F.12: Comparison of species trees built using FastTree and the robust strategy. The

"robust strategy" refers to RAxML + CAT for tree topology, and IQ-TREE + Gamma for branch-lengths. A series of taxon sets downsampled from the original 10,575 genomes (same as shown in Supplementary Fig. F.13) were tested. **a** and **b**. Distances between pairs of FastTree vs. robust trees on the same dataset. **c**. Distances among FastTree trees on different datasets. **d**. Distances among robust trees on different datasets. **e**. PCoA on RF distance matrix among all trees. **f**. PCoA on tip distance matrix among all tree. Pairs of FastTree (diamond) and robust (circle) trees on the same dataset are connected by a line. Source data are provided as a Source Data file.

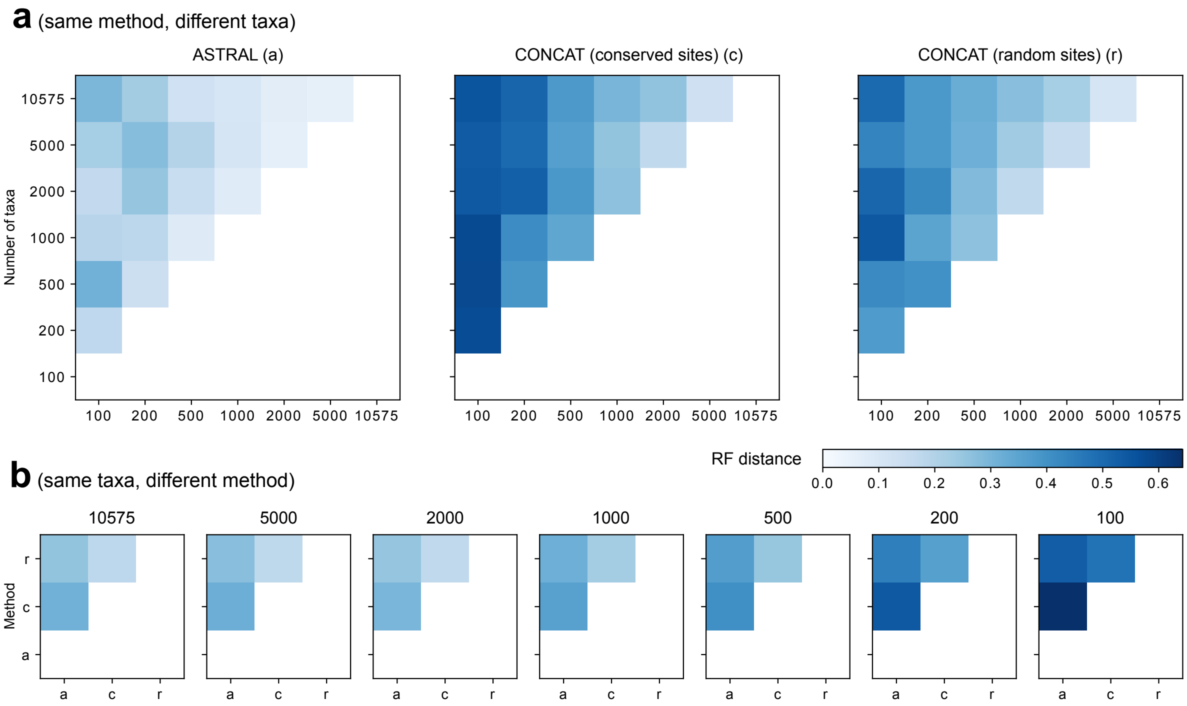


Figure F.13: Comparison of species trees built on a series of downsampled taxa. The original 10,575 taxa were subsampled to retain given number (5,000, 2,000, 1,000, 500, 200, and 100) of taxa representative of deep, large clades, as determined using the RED metric (see Methods). Three methods: ASTRAL, CONCAT (using most conserved or randomly selected sites) were evaluated. **a**. RF distance matrices of trees among taxon sets and within each method. **b**. RF distance matrices of trees across methods and within each taxon set. Source data are provided as a Source Data file.

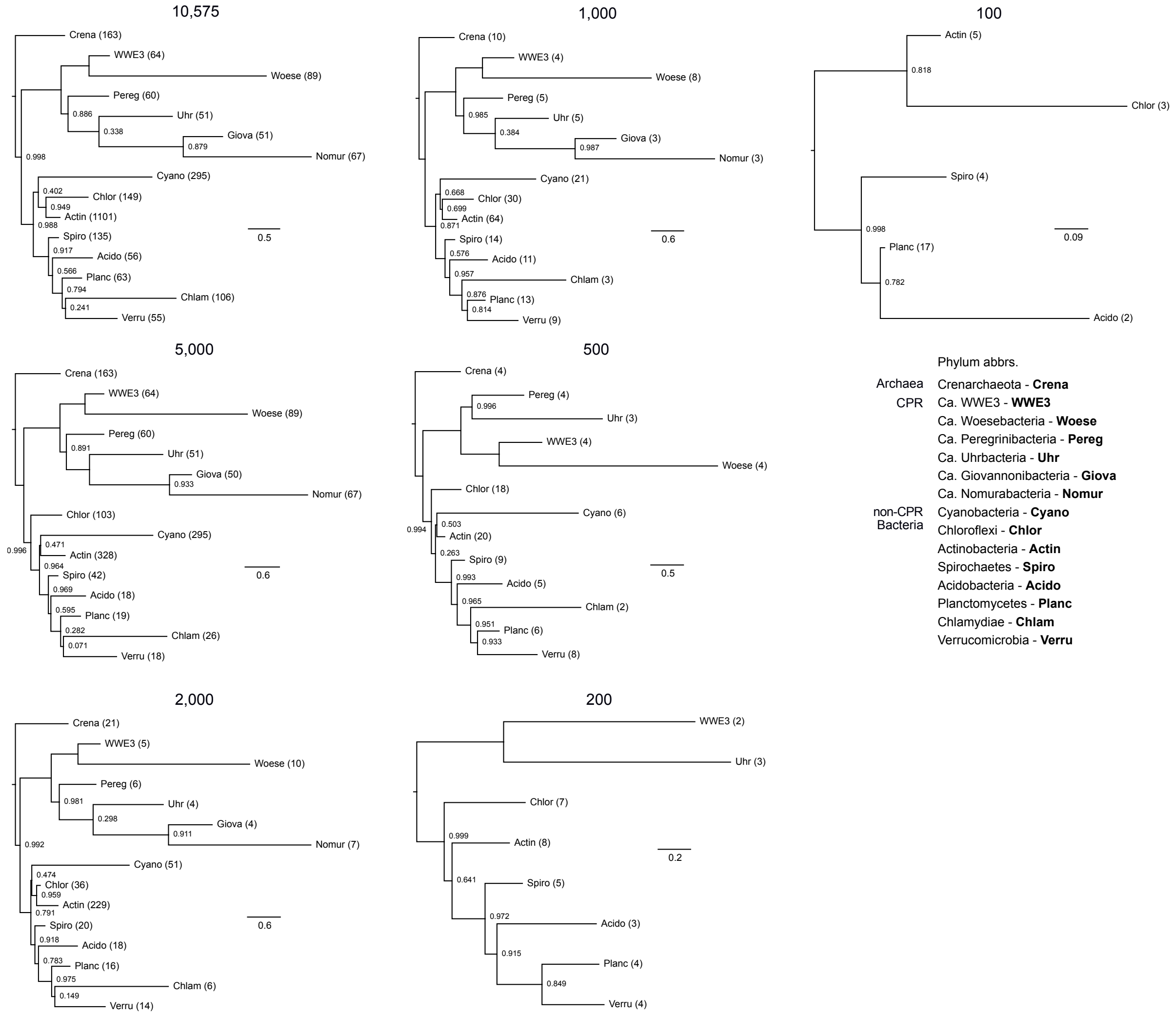


Figure F.14: Phylum-level relationships revealed by ASTRAL trees built on series of downsampled taxon sets. Panel headers indicate number of taxa. See Supplementary Fig. F.3's caption. Source data are provided as a Source Data file.

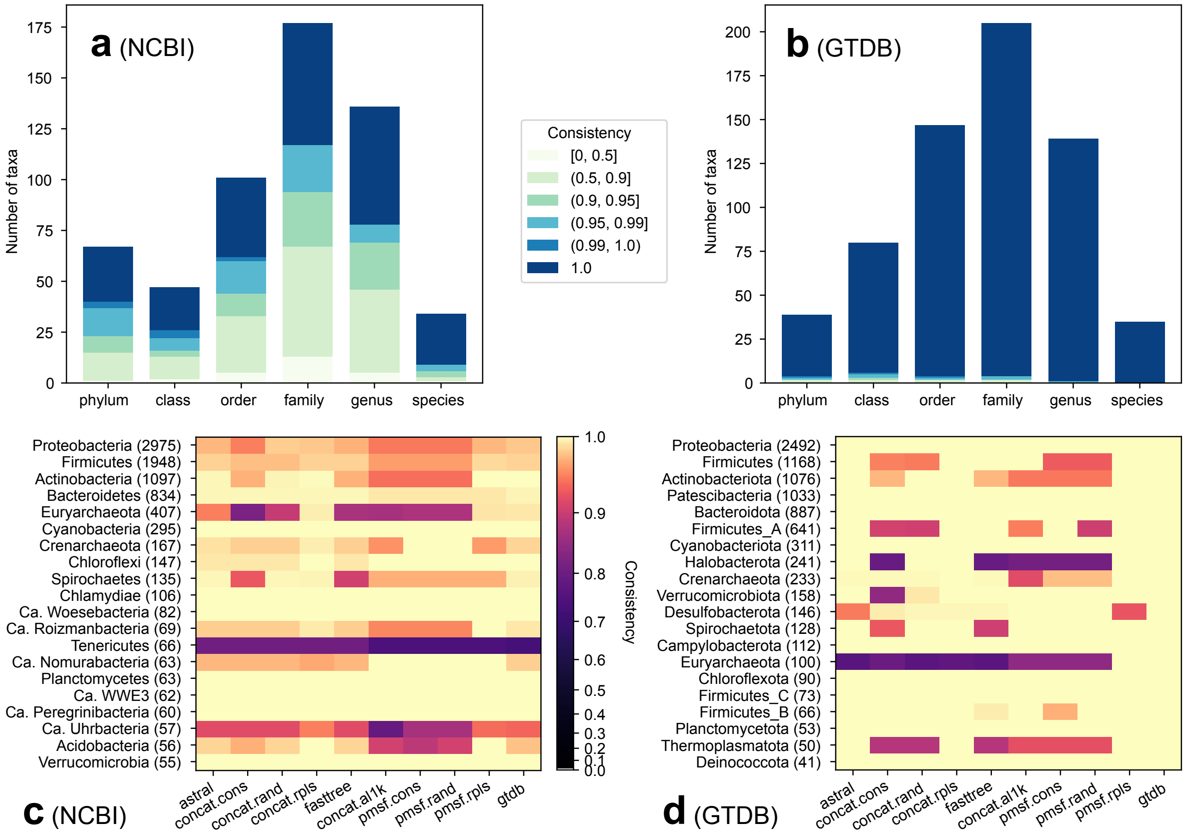


Figure F.15: Consistency of taxonomic units with phylogeny. Two taxonomy systems were evaluated: NCBI (**a** and **c**) and GTDB (**b** and **d**). The consistency scores were calculated using tax2tree (see Methods). **a** and **b**. Distribution of consistency scores of taxonomic units with at least ten representatives in the sampled genomes, calculated against the ASTRAL tree. **c** and **d**. Consistency scores of 20 most specious phyla of each system against each of the ten species trees (see Fig. 6.3). Numbers in parentheses indicate the number of taxa assigned to each group by tax2tree against the ASTRAL tree. Source data are provided as a Source Data file.

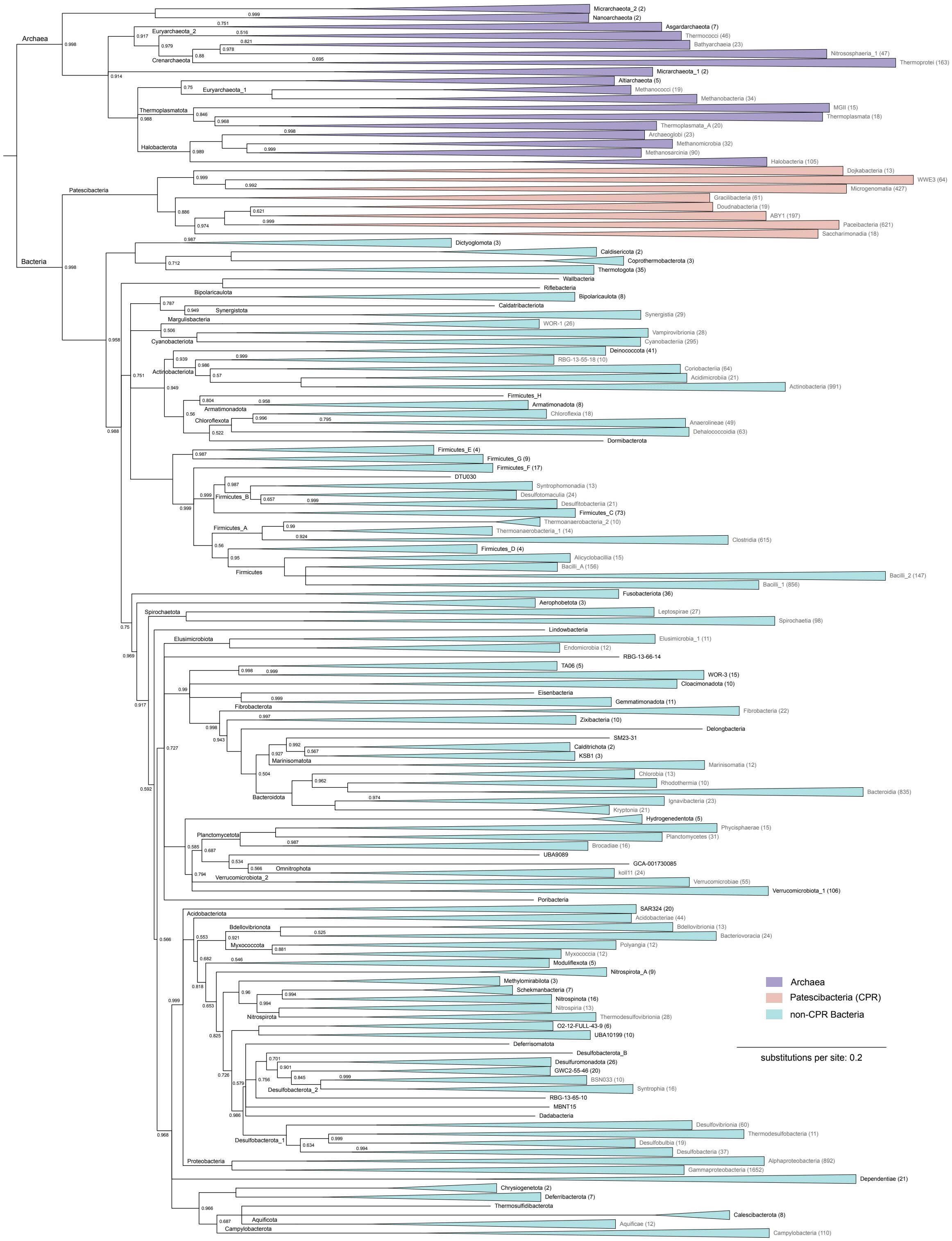


Figure F.16: The ASTRAL summary tree annotated using the GTDB taxonomy, collapsed to class level. The tree is identical to that in Supplementary Fig. F.4, except for the taxonomic annotations and the alternative collapsing pattern based on taxonomy. The three major groups

discussed in this study: Archaea, CPR and non-CPR Bacteria, were colored following Fig. 6.4a, b. But note that in GTDB, CPR is classified as phylum Patescibacteria. The triangles represent collapsed clades, with length equal to the longest branch in the clade. Node labels represent local posterior probability (lpp) of the corresponding branch. Labels are omitted at fully-supported (lpp = 1.0) branches. Source data are provided as a Source Data file.

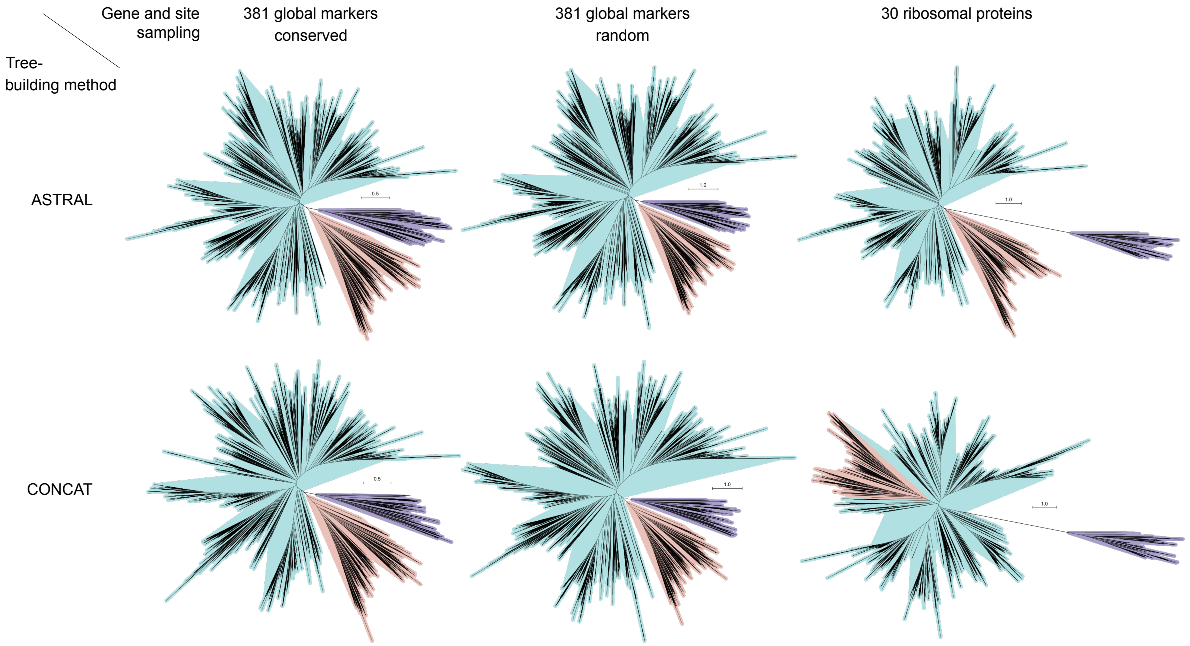


Figure F.17: Dimensions and separation of domains Archaea and Bacteria. This extends Fig. 6.4a, b (with the same color code) to all six trees using different methods (ASTRAL or CONCAT), gene sampling (381 global markers or 30 ribosomal proteins) and site sampling (most conserved or randomly selected). The three top panels are the same topology (the ASTRAL tree), with branch lengths re-estimated using different concatenated alignments. The three bottom panels are different trees separately reconstructed using the corresponding concatenated alignments. Note that in the CONCAT tree by ribosomal proteins, the placement of CPR could not be resolved, thus not depicted as a sister group to non-CPR Bacteria. All trees were drawn to scale, without collapsing or downsampling. Source data are provided as a Source Data file.

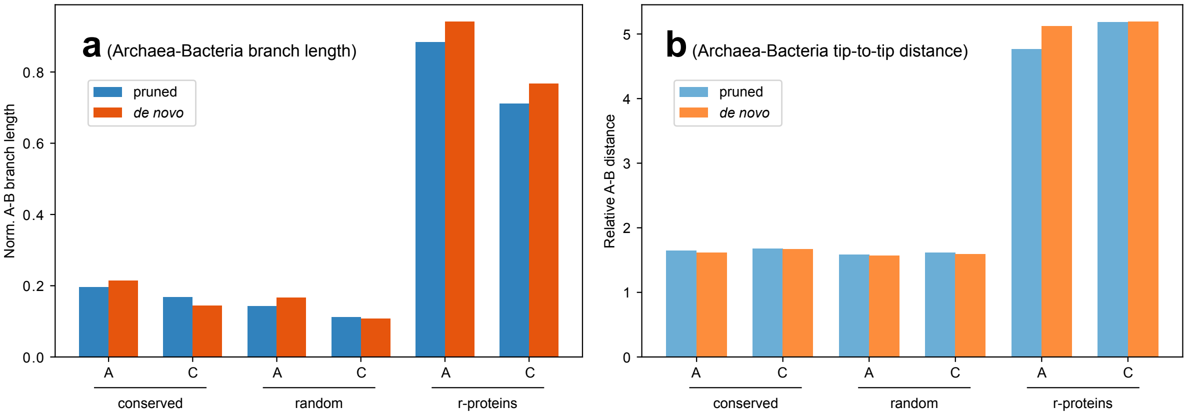


Figure F.18: Domain-level phylogenetic distances indicated by trees without CPR taxa. The normalized Archaea-Bacteria branch length (a) and the relative Archaea-Bacteria distance (b) (see Fig. 6.4e, f) of each tree are shown. “Pruned” are the same trees from the main results (Fig. 6.4e, f), with the CPR clade pruned; “*de novo*” are trees reconstructed from CPR-free sequence alignments. Each group contains six trees, built using either (A)STRAL or (C)ONCAT, with either conserved or random site sampling from the 381 global markers, or with the 30 ribosomal proteins. Source data are provided as a Source Data file.

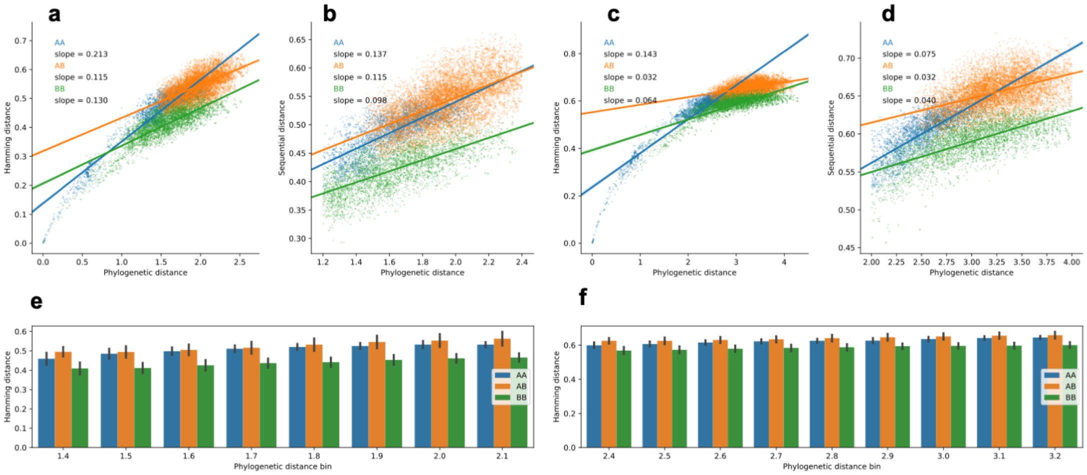


Figure F.19: Test for amino acid substitution saturation using conserved or random sites. The pairwise phylogenetic distances (sum of branch lengths) among 100 randomly sampled genomes from each domain are plotted. AA and BB represent intra-domain (Archaea-Archaea and Bacteria-Bacteria, respectively) distances while AB represents inter-domain (Archaea-Bacteria) distances. **a-d**: Scatter plots of Hamming distances determined based on pairwise sequence alignments vs. phylogenetic distances. Linear regression lines for the three groups are depicted respectively, with their slopes annotated. **e-f**: Phylogenetic distances were binned at equal intervals where each group has a sample size of five or larger. Error bars represent 95% confidence intervals computed from 1,000 bootstraps. The sequence alignments used for computing the Hamming distances were the most conserved sites for **a**, **b** and **e**, and the randomly selected sites for **c**, **d** and **f**. Panels **b** and **d** are zoom-in views of **a** and **c** to show the phylogenetic distance ranges where all three groups are populated. Source data are provided as a Source Data file.

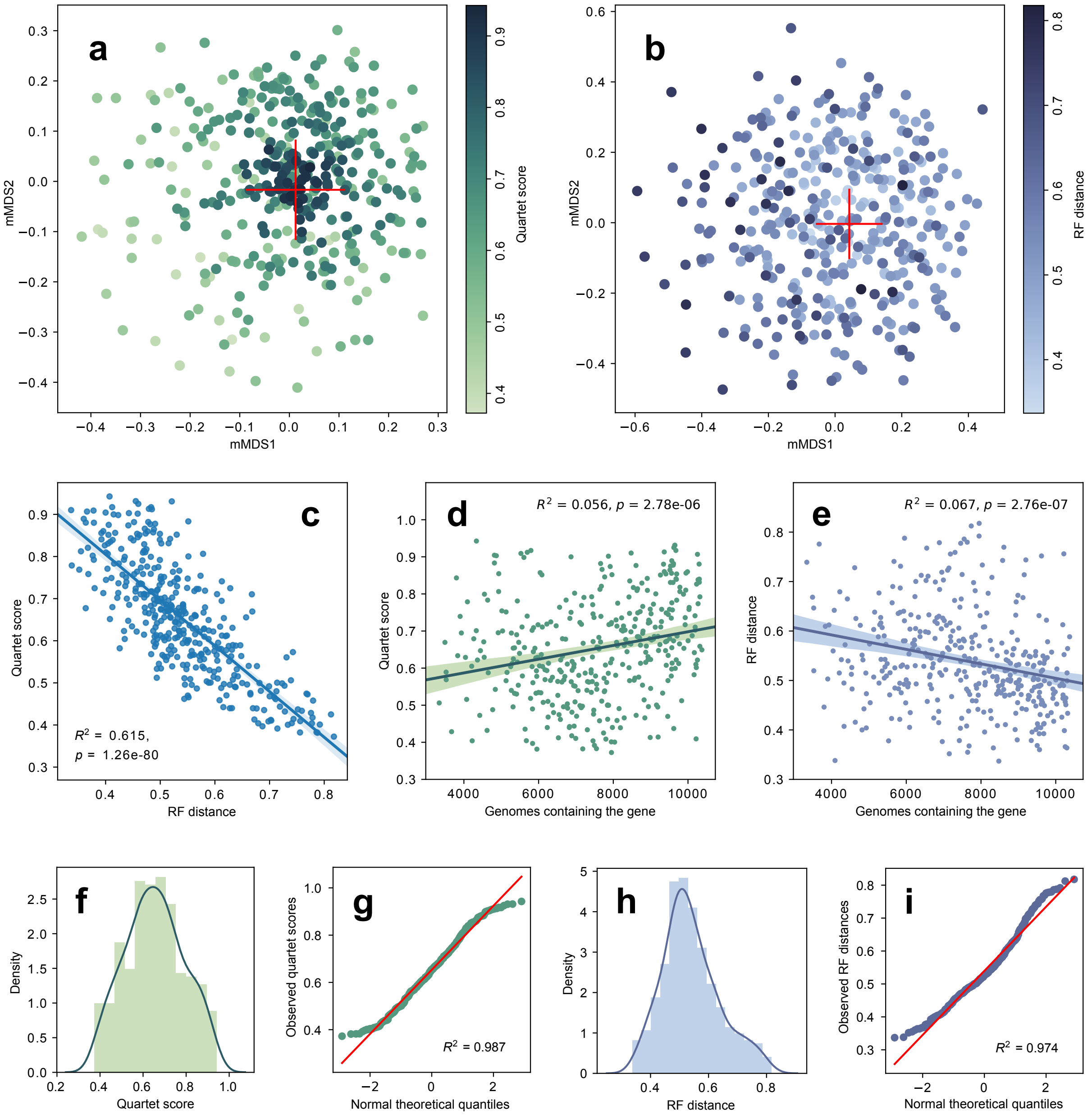


Figure F.20: Concordance among individual gene trees and the ASTRAL species tree. **a**. metric multidimensional scaling (mMDS) plot based on the quartet distance (1 - quartet score) between each pair of the 381 gene trees plus the species tree. The center of the red cross indicates the position of the species tree. **b**. mMDS plot based on the Robinson–Foulds (RF) distances. **c**. Linear regression between the quartet score and the RF distance. **d**. Linear regression between the quartet score and the number of genomes in which the corresponding gene was detected. **e**. Linear regression between the RF distance and the number of genomes in which the corresponding gene was detected. The squared Pearson correlation coefficient (R^2) and two-tailed p-value are displayed for each linear regression. **f**. Histogram and kernel density plot of the quartet scores of the 381 gene trees vs. the species tree. **g**. Quantile-quantile (Q-Q) plot

showing how well the quartet scores (*y*-axis) fit a normal distribution (*x*-axis). **h**. Histogram and kernel density plot of the RF distances of the gene trees vs. the species tree. **i**. QQ plot showing the fitness of the RF distances to a normal distribution. The coefficient of determination (R^2) is displayed for each Q-Q plot. Source data are provided as a Source Data file.

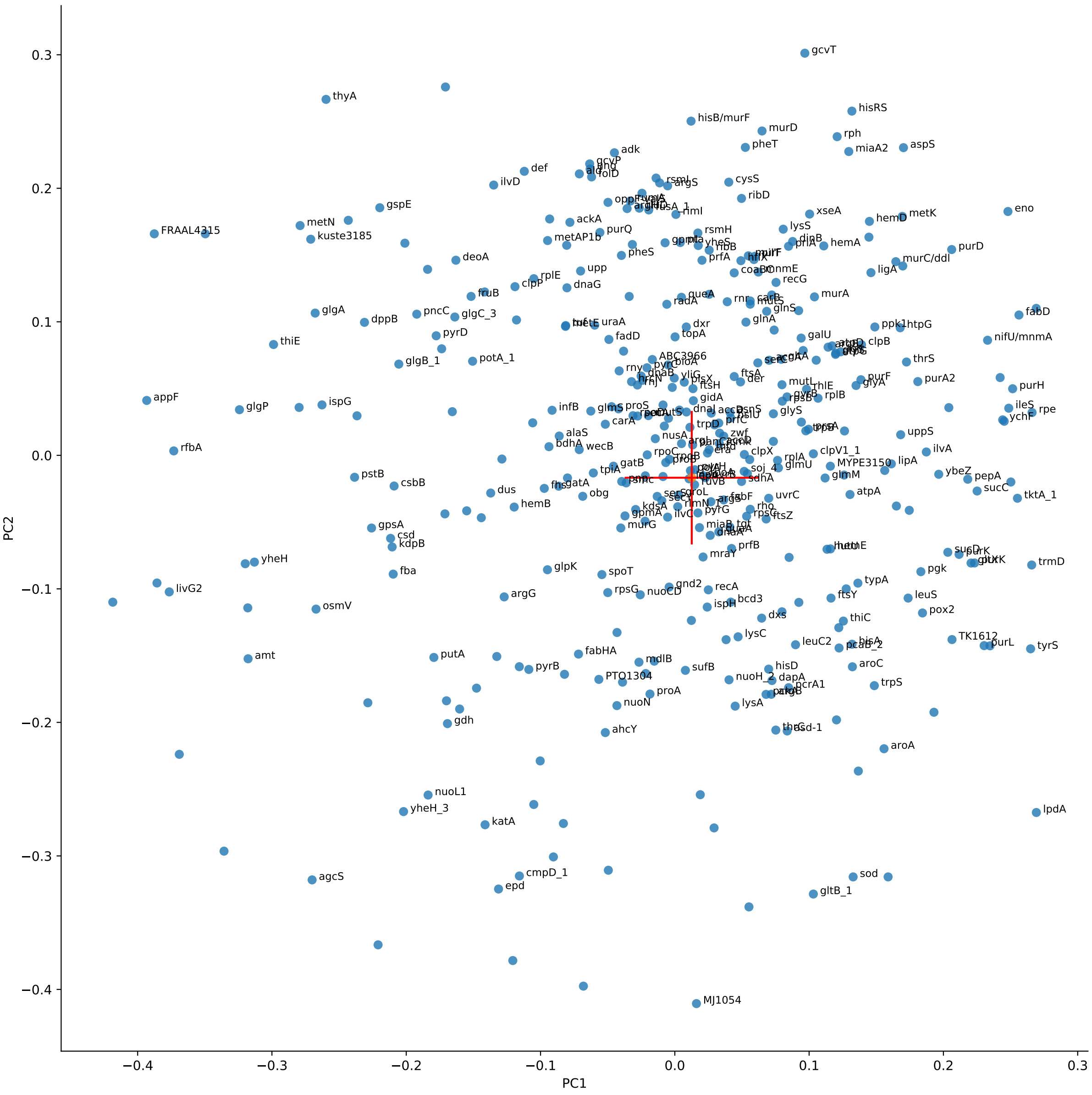


Figure F.21: mMDS plot by pairwise quartet distances among the 381 gene trees and the ASTRAL species tree. This is an enlarged view of Fig. 6.5d and Supplementary Fig. F.20a. If a marker gene was annotated with an official gene name from the UniProt database (see https://www.uniprot.org/help/gene_name for rules), the corresponding gene tree is labeled with that name. Source data are provided as a Source Data file.

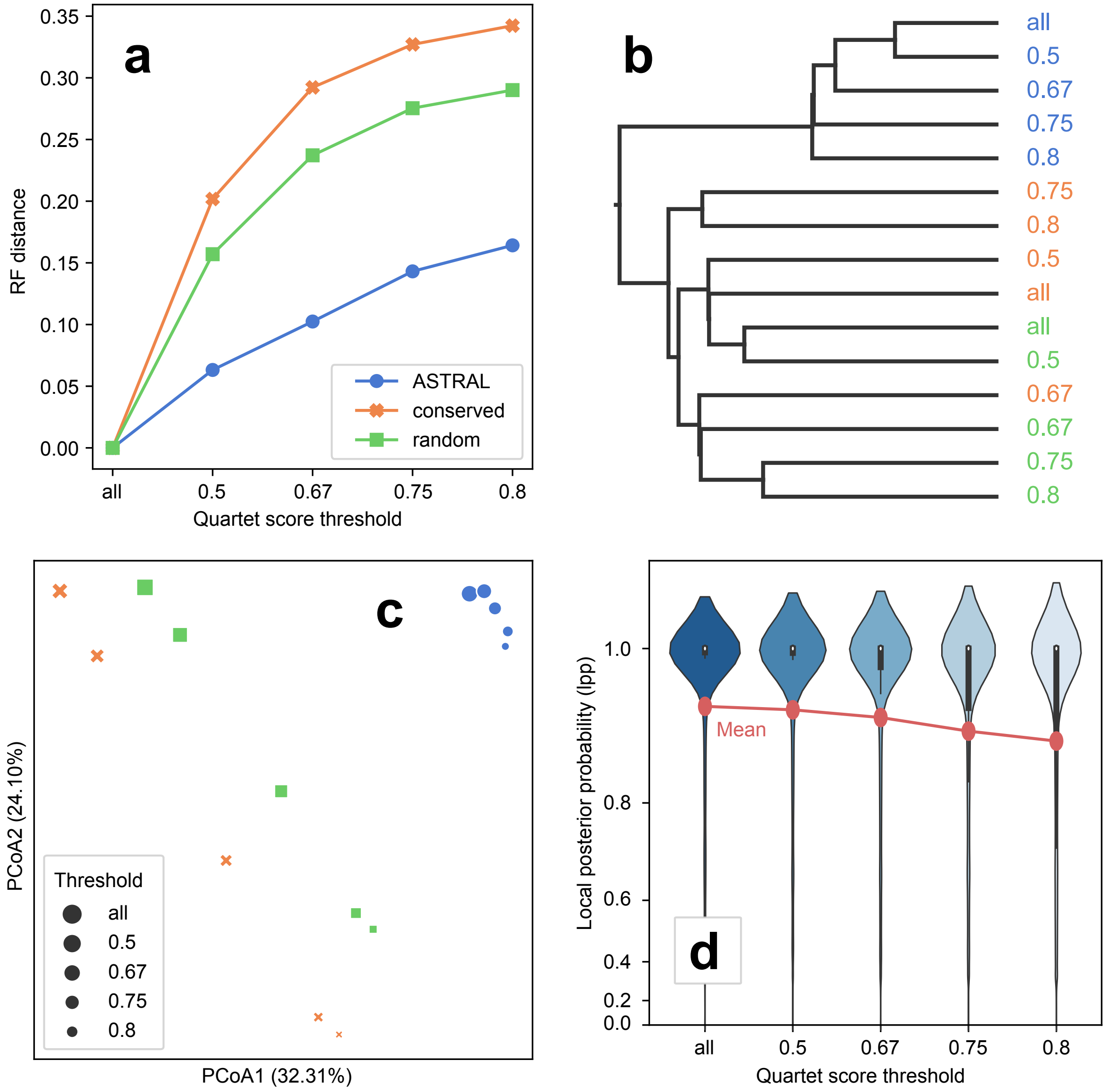


Figure F.22: Comparison of species trees built using marker genes subsampled by quartet score. The 381 marker genes (all) were downsampled to subsets in which the quartet score of the corresponding gene tree is at least 0.5 (322 genes), 0.67 (171 genes), 0.75 (93 genes) and 0.8 (64 genes), respectively. Three methods: ASTRAL (blue), CONCAT by most conserved sites (orange) or randomly selected sites (green) were tested. **a**-**c**: Topological discrepancy between pairs of trees, as measured by the Robinson–Foulds (RF) distance. **a**. RF distance from tree on each subset to the fullscale tree ("all") by method. **b**. Hierarchical clustering of the RF distance matrix. **c**. PCoA of the RF distance matrix. **d**. Violin plots of distribution of ASTRAL tree branch supports (lpp) on each subset. The red lines represent means. The *y*-axis is in exponential scale. Source data are provided as a Source Data file.

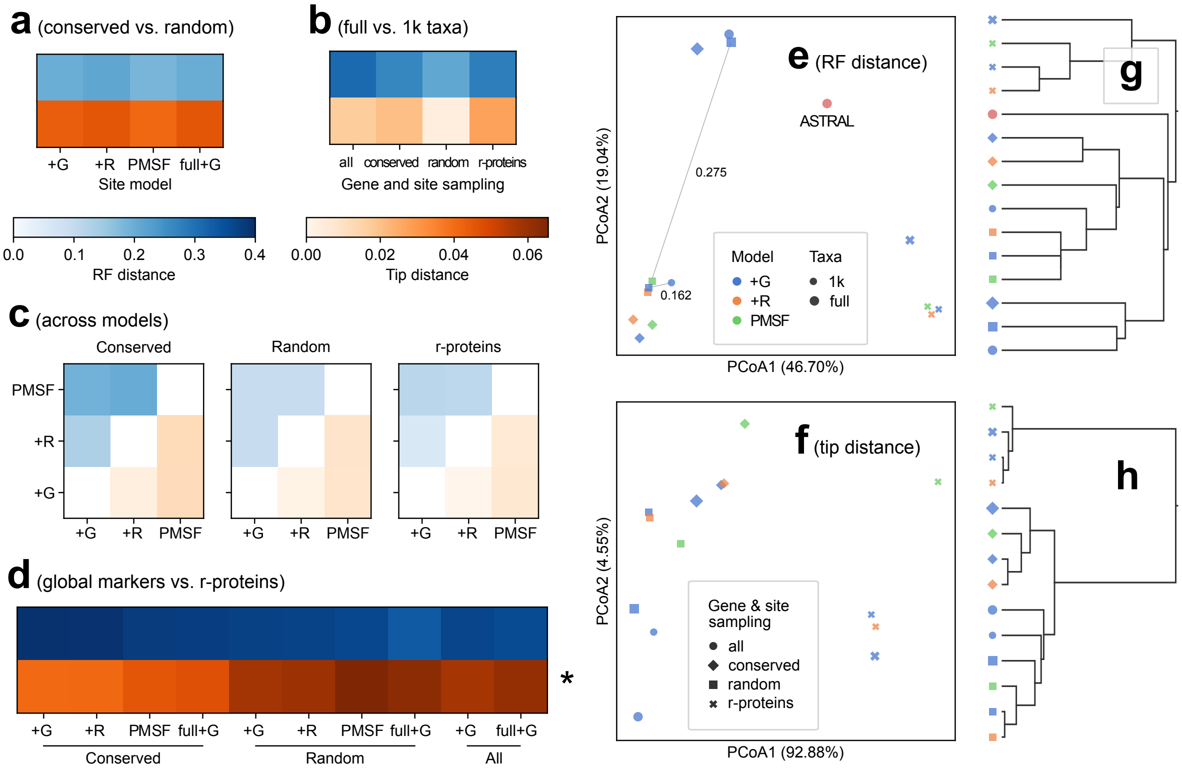


Figure F.23: Comparison of CONCAT trees on downsampled 1,000 taxa and alternative site models. The 1,000-taxon set is the same as shown in Supplementary Fig. F.13. Three models are compared: “+G”: the conventional Gamma model, i.e., the rate heterogeneity across sites is subject to a Gamma distribution; “+R”: the FreeRate model, which relaxes the assumption of Gamma distribution of rates; “PMSF”: the posterior mean site frequency model, which operates on site profiles determined by the profile mixture model C60 (selected in a model test). As controls, the 10,575-taxa full-scale CONCAT trees were truncated to the 1,000 taxa for comparison (“full+G”). Blue: RF distances. Orange: tip distances. **a**. Distances between trees by differential site sampling: most conserved or randomly selected sites. **b**. Distances between trees by differential taxon sampling: 10,575 (full) or 1,000 taxa, both using the Gamma model. **c**. Distances among trees by different site models. **d**. Distances among trees based on the 381 global marker genes or the 30 ribosonal proteins. Note (*) that the tip distances illustrated in this panel were divided by three, otherwise they would be too dark to allow other panels being distinguishable. **e**. PCoA of RF distance matrix. A special comparison between the impact of site sampling vs. that of taxon sampling was highlighted by grey lines, and the corresponding RF distances were annotated. **f**. PCoA of tip distance matrix. **g**. Hierarchical clustering of RF distance matrix. **h**. Hierarchical clustering of tip distance matrix. For **e** and **g**, the ASTRAL tree (red) was included as a reference, but it was not included in **f** and **h** because ASTRAL does not directly compute branch lengths in unit of substitutions per site. Source data are provided as a Source Data file.

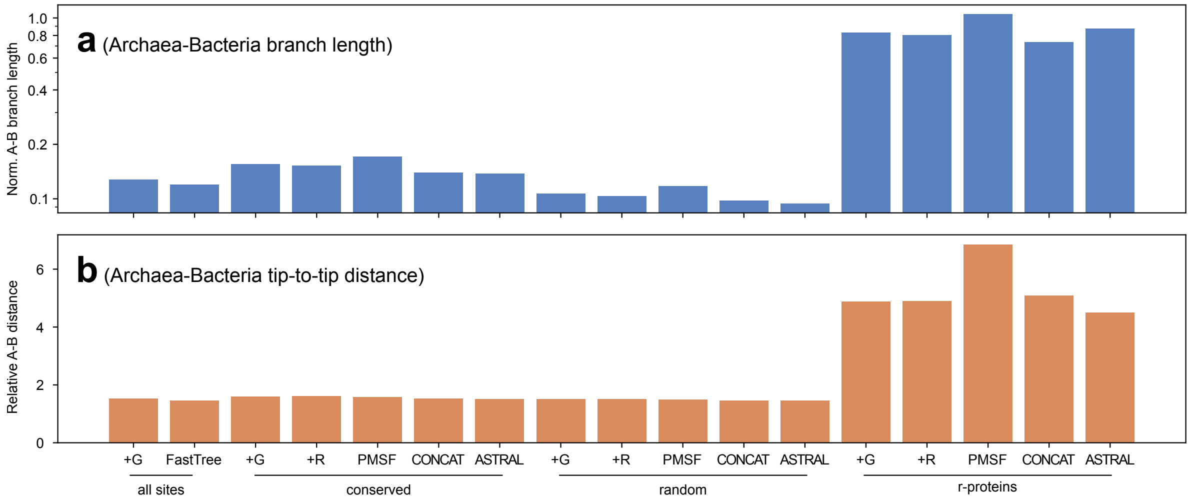


Figure F.24: Domain-level phylogenetic distances indicated by the 1,000 downsampled taxa. The normalized Archaea-Bacteria branch length (**a**) and the relative Archaea-Bacteria distance (**b**) (see Fig. 6.4f) of each tree are shown. Being compared are trees reconstructed based on the 1,000 taxa (+G, +R and PMSF), and trees inferred based on all 10,575 taxa but pruned to retain the same 1,000 taxa (FastTree, CONCAT and ASTRAL, branch lengths all based on the Gamma model). Note that these metrics are not directly comparable to those of the full-scale trees shown in Fig. 6.4e, f, due to taxon downsampling. Source data are provided as a Source Data file.

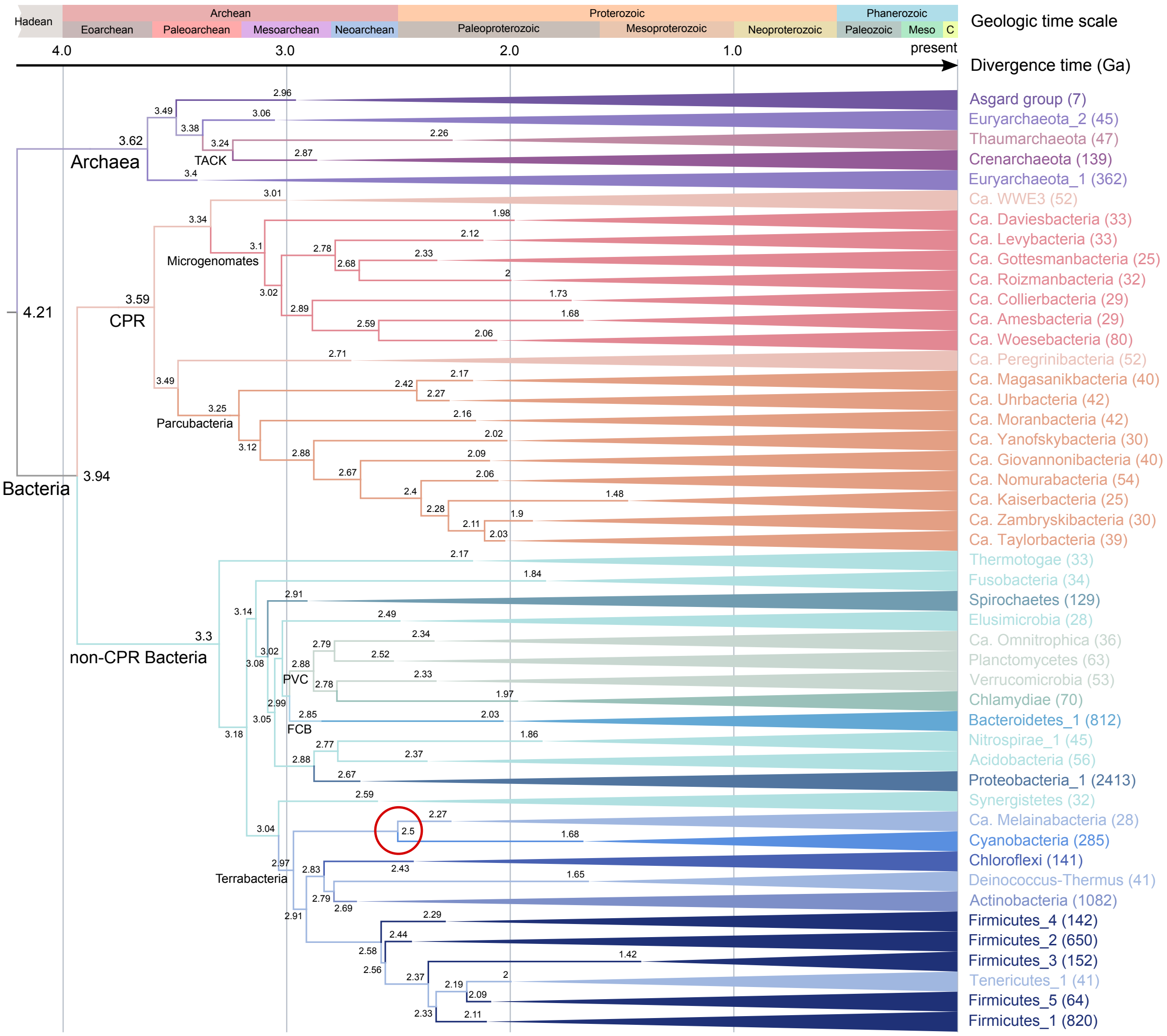


Figure F.25: Chronogram of microbial evolution inferred using maximum likelihood with a strict clock model. The evolutionary times were inferred based on the ASTRAL tree with branch lengths re-estimated using the conserved sites, and calibrated by the predicted emergence of the photosynthetic cyanobacteria (indicated by a red circle). For display purpose, clades representing phyla with at least 25 descendants were preserved and collapsed as triangles. Node labels represent the time in Ga (billion years ago) estimated by the run with the best likelihood score out of 10 replicates. The color scheme is consistent with Fig. 6.1. Source data are provided as a Source Data file.

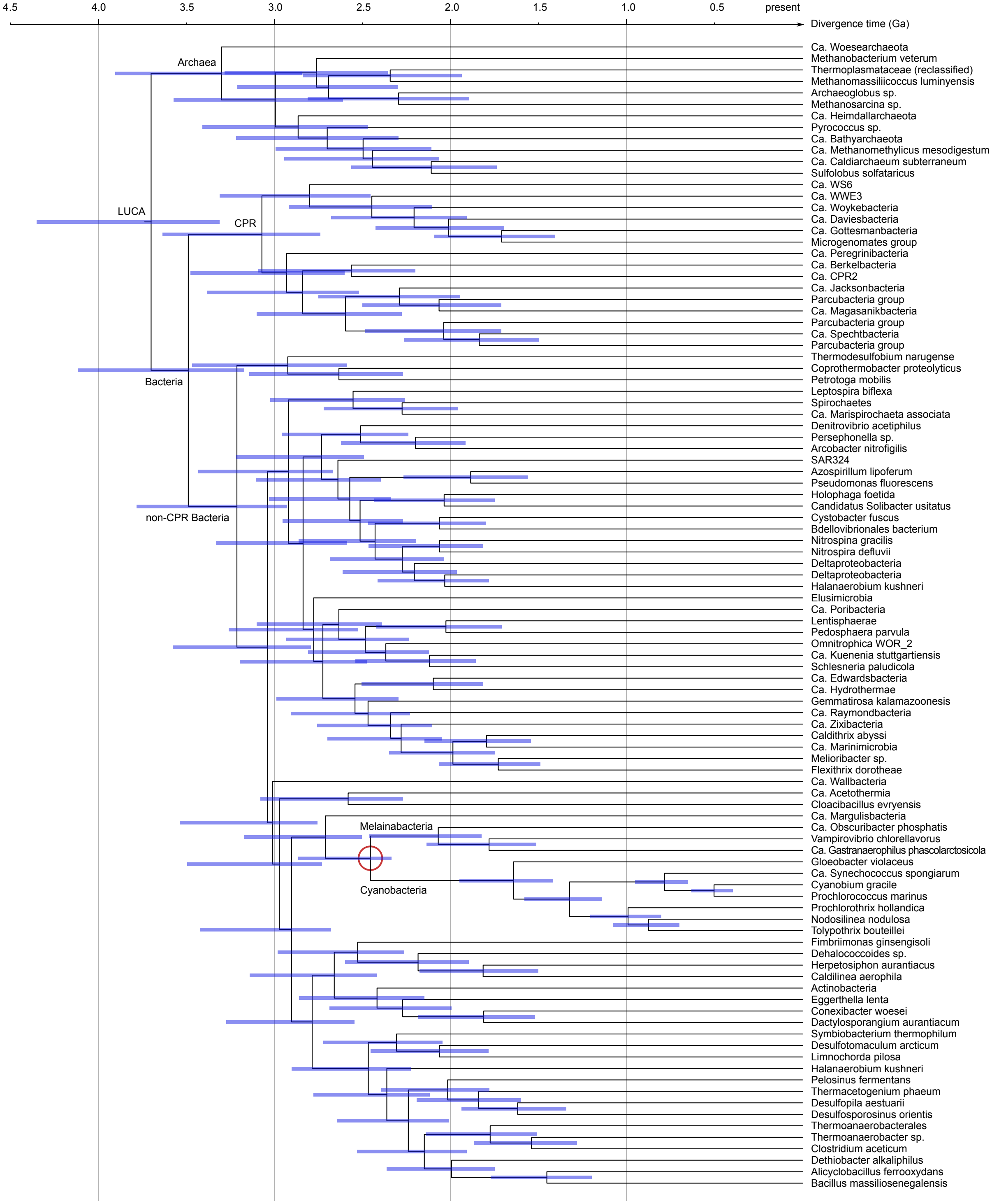


Figure F.26: Chronogram of microbial evolution inferred using Bayesian with a relaxed clock model. One hundred taxa by 5,000 randomly sampled sites were included in this analysis. The tree topology is identical to the ASTRAL tree. The node where time constraint (using the "wide" prior distribution) was placed on is indicated by a red circle. Node ages were estimated using BEAST, with an uncorrelated lognormal relaxed clock model (UCLD). Taxon labels are the

Latin species names, wherever available, omitting strain names, or the higher rank (usually phylum or superphylum) name if underclassified. Node heights represent the median of sampled age estimates of the node. Node bars indicate 95% confidence intervals. Source data are provided as a Source Data file.

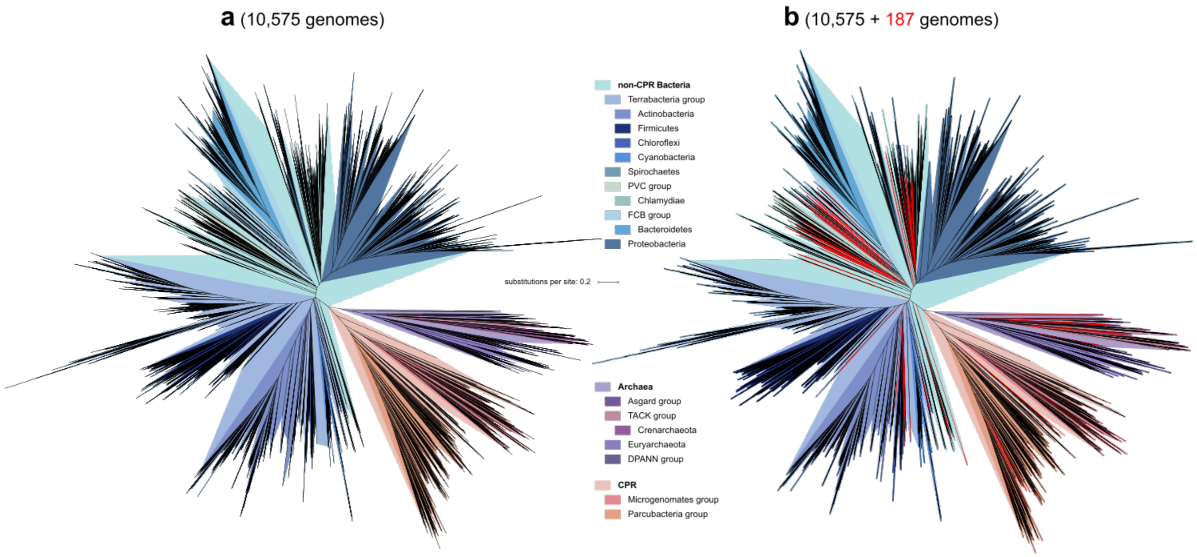


Figure F.27: Consistency of reconstructed evolutionary relationships with newly discovered microbial diversity. Both trees were built using ASTRAL on the 381 marker genes, and the branch lengths were estimated using up to 100 most conserved sites per gene. Color codes of clade shadows are consistent with Fig. 6.1. The trees are drawn-to-scale, with all taxa displayed. **a**. Tree of 10,575 genomes, which is the same as shown in Figs. 1, 3a and S5. **b**. Tree of the same 10,575 genomes plus 187 new genomes as of May 2019, representing previously missing or underrepresented NCBI and GTDB phyla. Clades constituted of the new genomes are colored red. Source data are provided as a Source Data file.

F.3 Supplementary Tables

Name	Date	Publication	Domain(s)	Phylogenetic tree			Character matrix			Related works
				Taxa	Gene(s)	Method	Taxa	Characters	Unit	
Woese and Fox	1977-11-01	270744	A, B, E	13	SSU	“comparative”	N/A	N/A	bp	2112744
Barns et al.	1996-08-20	8799176	A, B, E	64	SSU	fastDNAml	64	N/A	bp	9115194
Ciccarelli et al.	2006-03-03	16513982	A, B, E	191	31	PhyML	181	999,326	aa	
LTP rel. 93	2008-08-09	18692976	A, B	6,727	SSU	RAxML	9,975	14,576,220	bp	
AMPHORA	2008-10-13	18851752	B	578	31	PhyML	578	4,033,260	aa	20033048
Cox et al.	2008-12-23	19073919	A, B, E	40	45	P4	40	N/A	aa	24336283
Greengenes rel. 13_5	2013-05-20	22134646	A, B	203,452	SSU	FastTree	203,452	260,068,849	bp	
Lang et al.	2013-04-25	23638103	A, B	841	24	BUCKy	840	3,601,341	aa	
GEBA-MDM	2013-07-14	23851394	B	2,229	38	RAxML	2,228	16,304,266	aa	
PhyloPhlAn	2013-08-14	23942190	A, B	3,737	400	RAxML	3,139	10,399,954	aa	
Hug et al.	2016-04-11	27572647	A, B, E	3,083	16	RAxML	3,080	6,532,247	aa	29522741
1,003 GEBA genomes	2017-06-12	28604660	A, B	1,003	56	RAxML	1,039	17,750,144	aa	
Schulz et al.	2017-10-17	29041958	B	12,400	SSU	RAxML	926	1,343,426	bp	
GTDB rel. 80	2018-08-27	30148503	B	21,943	120	FastTree	21,547	650,103,222	aa	28894102
this work	N/A	N/A	A, B	10,575	381	RAxML	10,474	273,417,890	aa	
							10,485	265,218,697	aa	
						ASTRAL	10,575	1,162,421,084	aa	

Table F.1: A summary of previous and current trees of microbial life.

The table summarizes representative phylogenetics studies that featured the global taxon sampling of one or multiple domains of microbial life forms. Only works involving de novo phylogenetic reconstructions based on the entire datasets were selected (thus excluding synthesis studies such as the Open Tree of Life [132]). The name of each work is either the project name plus the release version, if applicable, or in the “authors (year)” format. “Date” is the date of the release (if applicable) or the publication. “Publication” is the NCBI PMID of the article. For one work containing multiple trees, only one tree that was based on the largest dataset, built using the most expensive method, or recommended by the authors was recorded. For one series of closely related works, only one work that was most relevant in the context of “tree of life” was recorded, while the others were mentioned in the “related works” field. “Domain(s)” codes are (A)rchaea, (B)acteria and (E)ukaryota. In some works (such as GEBA-MDM and GTDB), because taxa from different domains were subjected to separate phylogenetic reconstructions, only the largest domain (Bacteria) was recorded. Whenever possible, the actual dimensions of the phylogenetic tree and the supporting character matrix (i.e., a multiple sequence alignment, excluding duplicates) were recorded. “Characters” is the sum of non-missing, non-gap characters (unit: bp (basepair) or aa (amino acid)). Note that the numbers of taxa in the tree and in the matrix may be different due to filtering and clustering operations.

1. Pre-tree-building steps	Marker extraction (x10575) (PhyloPhlAn)		Alignment (x381) (PASTA / UPP)		Model selection (x381) (RAxML)		Total		
	runtime (hr)	CPU hrs (x32)	runtime (hr)	CPU hrs (x3)	runtime (hr)	CPU hrs (x4)	CPU hrs		
	17.78	568.96	772.12	2316.36	932.90	3731.60	6616.92		
2. Tree-building-summary	Gene tree building (x381) (starting tree) (FastTree)		Gene tree building (x381) (RAxML / IQ-TREE)		Gene tree summarization (ASTRAL)			Total	
	runtime (hr)	CPU hrs (x4)	runtime (hr)	CPU hrs (x24)	runtime (hr)	CPU hrs (x28)	GPU hrs (x4)	CPU hrs	GPU hrs
	213.39	853.56	3980.87	95540.88	9.96	278.85	39.84	96673.29	39.84
3. Tree-building-concatenation	Starting tree building (FastTree)		Tree topology search (RAxML + CAT)		Tree optimization (IQ-TREE + Gamma)		Rapid bootstrap (x100) (RAxML + CAT)		Total
site sampling	runtime (hr)	CPU hrs (x3)	runtime (hr)	CPU hrs (x24)	runtime (hr)	CPU hrs (x24)	runtime (hr)	CPU hrs (x24)	CPU hrs
conserved	6.79	20.37	143.20	488.88	1.55	37.27	1362.58	32701.80	33248.32
random	7.03	21.09	156.45	506.16	1.37	32.93	1487.49	35699.69	36259.87

Table F.2: Computational expenses for building the phylogenies of 10,575 microbial genomes based on 381 marker genes.

For each procedure, the runtime (wall-clock time) is listed, and the charged time (CPU hours or GPU hours) was obtained by multiplying the runtime by the number of CPU cores or GPU units allocated (shown in parentheses). Times for procedures that were inexpensive or not directly relevant to the tree-building have been omitted. Several steps consisted of multiple independent jobs that can be effectively parallelized. The number of jobs is indicated in parentheses after the procedure title. Several steps actually consisted of multiple trials (e.g., we did three runs per maximum likelihood tree building and selected the one with the highest Gamma likelihood), but in this table we only report the times of the selected trials. Thus, this table indicates the minimum time required for building the phylogenies we present.

Rank	**Same**	**Add**	**Change**	**Delete**	**Empty**
phylum	10000	158	50	177	190
class	7900	304	90	90	2191
order	7625	439	183	104	2224
family	7074	423	299	251	2528
genus	6655	159	350	624	2787
species	10229	0	247	99	0

Table F.3: Summary of NCBI taxonomy curated based on phylogeny.

“Same”: validated the original assignment; “Add”: assigned a taxon to an originally unassigned rank; “Change”: modified an originally incorrectly assigned taxon; “Delete”: deleted an originally incorrectly assigned taxon; “Empty”: unassigned in both original and curated taxonomy.

Gene sampling	Site sampling	Method	Radius	A-B branch length	Norm A-B branch length	A depth	A-B branch/ A depth	B depth	A-B branch/ B depth	Mean A-A distance	Mean B-B distance	Mean A-B distance	Relative A-B distance
global	conserved	ASTRAL	0.971	0.122	0.126	0.9	0.136	0.91	0.134	1.527	1.604	1.957	1.563
global	conserved	CONCAT	0.992	0.126	0.127	0.873	0.144	0.932	0.135	1.514	1.639	1.99	1.596
global	random	ASTRAL	1.773	0.152	0.086	1.405	0.108	1.717	0.089	2.343	3.015	3.274	1.517
global	random	CONCAT	1.801	0.159	0.088	1.436	0.111	1.739	0.091	2.356	3.079	3.35	1.547
r-proteins	all	ASTRAL	3.018	2.528	0.838	1.589	1.591	1.767	1.431	2.449	3.068	5.815	4.501
r-proteins	all	CONCAT	3.333	2.324	0.697	1.823	1.275	2.2	1.057	2.51	3.218	6.348	4.99
global	all	FastTree	1.941	0.21	0.108	1.393	0.151	1.858	0.113	2.509	3.233	3.522	1.529

Table F.4: Evolutionary proximity between Archaea and Bacteria by differential gene, site sampling and method.

Letters “A” and “B” refer to Archaea and Bacteria, respectively. Two metrics were assessed: the length of the branch connecting LCA of Archaea and LCA of Bacteria, either original or normalized by the tree radius (calculated as the median of root-to-tip distances of all taxa); and the relative Archaea-Bacteria distance, calculated as: mean(A-B)2 / (mean(A-A) × mean(B-B)), in which each distance is the sum of lengths of branches connecting one tip to another. In addition, the depths of the Archaea and Bacteria clades, calculated as the median of root-to-tip distances of all taxa in each clade, and the length of the branch connecting the two LCAs divided by the clade depth, are provided, to reflect the proximity between Archaea and Bacteria as compared to the dimension of each clade.

Gene sampling	Site sampling	Method	Radius	A-B branch length	Norm A-B branch length	A depth	A-B branch/ A depth	B depth	A-B branch/ B depth	Mean A-A distance	Mean B-B distance	Mean A-B distance	Relative A-B distance
CPR clade pruned from tree													
global	conserved	ASTRAL	0.923	0.181	0.196	0.9	0.201	0.827	0.219	1.527	1.497	1.937	1.642
global	conserved	CONCAT	0.958	0.16	0.167	0.873	0.184	0.878	0.183	1.514	1.53	1.972	1.679
global	random	ASTRAL	1.744	0.249	0.143	1.405	0.177	1.637	0.152	2.343	2.907	3.287	1.587
global	random	CONCAT	1.804	0.201	0.112	1.436	0.14	1.722	0.117	2.356	2.983	3.369	1.615
r-proteins	all	ASTRAL	2.966	2.623	0.884	1.589	1.65	1.656	1.584	2.449	2.867	5.783	4.764
r-proteins	all	CONCAT	3.272	2.324	0.71	1.823	1.275	2.134	1.089	2.51	3.023	6.271	5.183
de novo tree from CPR-free alignment													
global	conserved	ASTRAL	0.956	0.204	0.213	0.871	0.234	0.853	0.239	1.556	1.512	1.946	1.61
global	conserved	CONCAT	0.98	0.141	0.144	0.888	0.159	0.91	0.155	1.543	1.566	2.009	1.67
global	random	ASTRAL	1.795	0.298	0.166	1.361	0.219	1.671	0.178	2.381	2.931	3.307	1.567
global	random	CONCAT	1.827	0.197	0.108	1.436	0.137	1.747	0.113	2.394	3.031	3.401	1.593
r-proteins	all	ASTRAL	3.308	3.115	0.942	1.563	1.993	1.769	1.761	2.593	3.046	6.361	5.123
r-proteins	all	CONCAT	3.272	2.511	0.767	1.87	1.343	2.039	1.232	2.54	3.064	6.358	5.193

Table F.5: Evolutionary proximity between Archaea and Bacteria by differential gene, site sampling and method.

The results of two experimental groups are shown. Upper: The CPR clade was pruned from the trees discussed in Fig. 6.3 and Supplementary Table F.4. Lower: The CPR sequences were removed from the dataset, and trees were re-built. The definitions of column names follow Supplementary Table F.4.

Gene sampling	No. of genes	Site sampling	Radius	A-B branch length	Norm A-B branch length	A depth	A-B branch/ A depth	B depth	A-B branch/ B depth	Mean A-A distance	Mean B-B distance	Mean A-B distance	Relative A-B distance
qts >0.5	322	conserved	0.863	0.132	0.153	0.859	0.154	0.795	0.166	1.426	1.387	1.801	1.641
qts >0.5	322	random	1.788	0.184	0.103	1.474	0.125	1.713	0.108	2.4	2.969	3.349	1.574
qts >0.67	171	conserved	0.887	0.241	0.272	0.928	0.26	0.756	0.319	1.538	1.328	1.929	1.822
qts >0.67	171	random	1.905	0.349	0.183	1.621	0.215	1.742	0.2	2.606	3.029	3.667	1.704
qts >0.75	93	conserved	0.831	0.278	0.334	0.917	0.303	0.679	0.409	1.5	1.2	1.888	1.98
qts >0.75	93	random	2.029	0.48	0.237	1.708	0.281	1.793	0.268	2.713	3.127	3.952	1.841
qts >0.8	64	conserved	0.812	0.276	0.34	0.957	0.288	0.663	0.416	1.494	1.186	1.886	2.007
qts >0.8	64	random	2.02	0.47	0.233	1.803	0.261	1.784	0.264	2.804	3.112	4.006	1.839

Table F.6: Evolutionary proximity between Archaea and Bacteria by differential gene, site sampling and method.

The 381 gene trees were subsampled based on their quartet scores (qts) vs. the species tree. Larger qts indicates higher topological concordance. The definitions of column names follow Supplementary Table F.4.

Gene sampling	Site sampling	Site model	Radius	A-B branch length	Norm. A-B branch length	A depth	A-B branch/ A depth	B depth	A-B branch/ B depth	Mean A-A distance	Mean B-B distance	Mean A-B distance	Relative A-B distance
global	all	CONCAT	1.333	0.17	0.128	1.047	0.162	1.26	0.135	1.856	2.225	2.518	1.535
global	conserved	Gamma	0.706	0.11	0.155	0.669	0.164	0.65	0.169	1.171	1.162	1.475	1.598
global	conserved	FreeRate	0.615	0.094	0.153	0.59	0.159	0.566	0.166	1.02	1.012	1.29	1.614
global	conserved	PMSF	0.982	0.168	0.171	1.049	0.16	0.885	0.19	1.833	1.66	2.197	1.586
global	random	Gamma	1.323	0.141	0.107	1.043	0.135	1.264	0.112	1.852	2.212	2.483	1.505
global	random	FreeRate	1.441	0.149	0.104	1.147	0.13	1.379	0.108	2.012	2.419	2.714	1.514
global	random	PMSF	2.203	0.259	0.118	1.836	0.141	2.083	0.125	3.268	3.67	4.234	1.495
r-proteins	all	Gamma	2.079	1.719	0.827	1.201	1.432	1.22	1.409	1.716	2.092	4.184	4.874
r-proteins	all	FreeRate	1.939	1.554	0.802	1.119	1.389	1.163	1.337	1.583	1.923	3.86	4.894
r-proteins	all	PMSF	4.048	4.237	1.047	1.775	2.387	1.934	2.191	2.857	3.286	8.016	6.845

Table F.7: Evolutionary proximity between Archaea and Bacteria with 1,000 taxa.

The original 10,575 genomes were downsampled to 1,000 (see Methods), which allowed for phylogenetic reconstruction using the more expensive site heterogeneous model PMSF, as compared to the simpler site homogeneous models Gamma and FreeRate. The definitions of column names follow Supplementary Table F.4.

Gene & sites	**Method**	**Reps. passed**	**LUCA**	**CPR split from Bacteria**	**Archea diversification**	**Non CPR Bacteria diversification**	**CPR diversification**
General results							
conserved	ASTRAL	9	4.228 ± 0.046 (4.206)	3.958 ± 0.043 (3.937)	3.9 ± 0.043 (3.879)	3.32 ± 0.036 (3.302)	3.772 ± 0.041 (3.752)
	CONCAT	8	4.181 ± 0.063 (4.147)	3.894 ± 0.058 (3.862)	3.855 ± 0.058 (3.824)	3.398 ± 0.051 (3.371)	3.736 ± 0.056 (3.705)
random	ASTRAL	8	3.631 ± 0.054 (3.618)	3.494 ± 0.052 (3.482)	3.295 ± 0.049 (3.284)	3.229 ± 0.048 (3.218)	3.276 ± 0.049 (3.265)
	CONCAT	7	3.654 ± 0.021 (3.7)	3.515 ± 0.02 (3.56)	3.28 ± 0.019 (3.321)	3.419 ± 0.02 (3.463)	3.299 ± 0.019 (3.341)
r-proteins	ASTRAL	10	7.068 ± 0.113 (7.174)	4.053 ± 0.065 (4.113)	3.470 ± 0.056 (3.522)	3.542 ± 0.057 (3.595)	3.945 ± 0.063 (4.004)
	CONCAT	9	7.012 ± 0.120 (6.963)	4.219 ± 0.072 (4.185)	3.689 ± 0.063 (3.659)	-	3.441 ± 0.059 (3.413)
Moving root on ASTRAL tree							
conserved	25%	7	4.211 ± 0.01 (4.213)	3.881 ± 0.009 (3.882)	3.986 ± 0.01 (3.987)	3.3 ± 0.008 (3.301)	3.716 ± 0.009 (3.718)
	75%	10	4.218 ± 0.066 (4.185)	4.043 ± 0.064 (4.013)	3.829 ± 0.06 (3.8)	3.337 ± 0.052 (3.311)	3.831 ± 0.06 (3.802)
random	25%	8	3.635 ± 0.058 (3.598)	3.456 ± 0.055 (3.421)	3.379 ± 0.054 (3.345)	3.216 ± 0.051 (3.184)	3.252 ± 0.052 (3.219)
	75%	10	3.589 ± 0.035 (3.568)	3.21 ± 0.031 (3.191)	3.21 ± 0.031 (3.191)	3.22 ± 0.031 (3.201)	3.279 ± 0.032 (3.26)
r-proteins	25%	10	7.131 ± 0.109 (7.066)	3.936 ± 0.06 (3.9)	3.638 ± 0.056 (3.605)	3.507 ± 0.054 (3.475)	3.848 ± 0.059 (3.813)
	75%	9	6.87 ± 0.056 (6.853)	4.186 ± 0.034 (4.177)	3.328 ± 0.027 (3.32)	3.558 ± 0.029 (3.55)	4.044 ± 0.033 (4.035)
PMSF vs. Gamma on 1k taxa							
all	Gamma	10	3.744 ± 0.017 (3.74)	3.525 ± 0.016 (3.521)	3.398 ± 0.016 (3.393)	3.271 ± 0.015 (3.267)	3.168 ± 0.015 (3.164)
conserved	Gamma	10	4.503 ± 0.02 (4.517)	4.156 ± 0.018 (4.168)	4.165 ± 0.019 (4.178)	3.509 ± 0.016 (3.52)	3.841 ± 0.017 (3.853)
	PMSF	10	4.553 ± 0.02 (4.543)	4.158 ± 0.019 (4.148)	4.265 ± 0.019 (4.255)	3.271 ± 0.015 (3.264)	3.869 ± 0.017 (3.86)
random	Gamma	10	3.718 ± 0.015 (3.712)	3.541 ± 0.014 (3.536)	3.419 ± 0.013 (3.414)	3.253 ± 0.013 (3.248)	3.205 ± 0.013 (3.2)
	PMSF	10	3.682 ± 0.015 (3.673)	3.486 ± 0.015 (3.477)	3.408 ± 0.014 (3.399)	3.192 ± 0.013 (3.185)	3.166 ± 0.013 (3.158)
r-proteins	Gamma	10	7.487 ± 0.03 (7.463)	4.416 ± 0.018 (4.402)	4.038 ± 0.016 (4.025)	4.224 ± 0.017 (4.211)	4.218 ± 0.017 (4.204)
	PMSF	10	9.219 ± 0.034 (9.19)	4.565 ± 0.017 (4.55)	3.939 ± 0.014 (3.926)	-	4.349 ± 0.016 (4.336)

Table F.8: Divergence time estimation results by maximum likelihood using one calibration.

The Cyanobacteria/Melainabacteria split was constrained to 2.5-2.6 Ga. For each setting, ten replicates were executed and the number of replicates that passed the gradient check was reported, and the means and standard deviations were calculated based on those replicates. The run with the best likelihood in all replicates was reported separately in parentheses. Estimated ages of five early evolutionary events were reported. The “non-CPR Bacteria diversification” field was left blank if the corresponding tree topology did not support the monophyly of non-CPR Bacteria.

Site sampling			conserved			random		
Name	Node	Range	Pass	LUCA	Non-CPR Bacteria diversification	Pass	LUCA	Non-CPR Bacteria diversification
Photosynthetic eukaryotes	Cyanobacteria LCA	>1.03	7	4.252 ± 0.076 (4.355)	3.338 ± 0.059 (3.419)	8	3.639 ± 0.06 (3.73)	3.237 ± 0.053 (3.318)
	Alphaproteobacteria LCA		9	4.232 ± 0.055 (4.201)	3.323 ± 0.043 (3.298)	6	3.589 ± 0.004 (3.587)	3.193 ± 0.004 (3.191)
	Alphaproteobacteria origin		9	4.242 ± 0.065 (4.201)	3.33 ± 0.051 (3.298)	9	3.625 ± 0.059 (3.728)	3.225 ± 0.052 (3.316)
	Cyanobacteria LCA and Alphaproteobacteria LCA		9	4.25 ± 0.068 (4.203)	3.337 ± 0.053 (3.3)	7	3.664 ± 0.071 (3.713)	3.259 ± 0.063 (3.303)
	Cyanobacteria LCA and Alphaproteobacteria origin		10	4.228 ± 0.054 (4.215)	3.32 ± 0.042 (3.309)	10	3.618 ± 0.058 (3.73)	3.218 ± 0.052 (3.318)
Akinetes-forming cyanobacteria	Nostocales origin	>1.2	10	5.534 ± 0 (5.534)	4.339 ± 0 (4.339)	7	4.695 ± 0 (4.695)	4.169 ± 0 (4.169)
		>1.5	10	6.253 ± 0 (6.253)	4.903 ± 0 (4.903)	7	5.302 ± 0 (5.302)	4.707 ± 0 (4.707)
		>1.9	10	7.163 ± 0 (7.163)	5.615 ± 0 (5.615)	9	6.084 ± 0 (6.084)	5.401 ± 0 (5.401)
		>2.1	10	7.594 ± 0 (7.594)	5.953 ± 0 (5.953)	7	6.459 ± 0 (6.459)	5.735 ± 0 (5.735)
Aphid-Buchnera symbiosis	Buchnera/Wigglesworthia split	>0.084	6	4.202 ± 0.002 (4.201)	3.299 ± 0.002 (3.298)	8	3.622 ± 0.057 (3.73)	3.222 ± 0.051 (3.318)
		>0.164	7	4.301 ± 0.076 (4.203)	3.376 ± 0.059 (3.299)	8	3.606 ± 0.05 (3.59)	3.208 ± 0.044 (3.193)
Photosynthetic eukaryotes and Aphid-Buchnera symbiosis	Cyanobacteria LCA and Alphaproteobacteria origin	>1.03	9	4.241 ± 0.058 (4.297)	3.33 ± 0.045 (3.373)	8	3.593 ± 0.011 (3.588)	3.196 ± 0.01 (3.191)
	Buchnera/Wigglesworthia split	>0.164						

Table F.9: Divergence time estimation results by maximum likelihood using alternative calibrations.

One or more calibrations were included in addition to the Cyanobacteria/Melainabacteria calibration, as described in each row. The definitions of column names follow Supplementary Table F.8.

Genes	Constraint	Prior dist.	Clock model	States (M)	MCMC Burn-in (M)	ESS	Age of LUCA (Ga)				Clock rate		CV	
							mean	median	95% low	95% high	ESS	mean	ESS	mean
global	narrow	norm	strict	10	1	1100	3.759	3.759	3.627	3.889	1337	0.288	-	-
	narrow	norm	ucld	50	5	431	3.821	3.816	3.56	4.089	526	0.289	449	0.176
	wide	ln	strict	10	1	5666	3.71	3.625	3.379	4.264	6996	0.293	-	-
	wide	ln	ucld	20	2	1007	3.768	3.7	3.312	4.351	1320	0.295	173	0.175
r-proteins	narrow	norm	strict	10	1	1390	7.45	7.448	7.127	7.765	1230	0.22	-	-
	narrow	norm	ucld	100	10	283	7.389	7.35	6.08	8.782	206	0.226	171	0.254
	wide	ln	strict	10	1	7645	7.347	7.198	6.64	8.455	7249	0.224	-	-
	wide	ln	ucld	50	10	250	7.362	7.254	5.782	9.142	296	0.229	157	0.255

Table F.10: Divergence time estimation results by Bayesian inference.

Input data were 100 taxa and 5,000 randomly sampled amino acid sites. Comparative analysis was performed using two clock models: strict clock or uncorrelated lognormal relaxed clock (ucld); two prior distributions of the time constraint of the Cyanobacteria/Melainabacteria split: “narrow”: a normal distribution which is narrower and based on previous estimates, and “wide”: a lognormal distribution which is wider and based on palaeobiological and geological evidence. We reported the estimated age of LUCA, the clock rate and its coefficient of variance (C.V., only for the relaxed clock model), which is a measurement of the “clock-likeness” of data (smaller is better).

Gene	Site			A-B	Norm. A-B	A	A-B	B	A-B	Mean	Mean	Mean	Relative
		Method	Radius	branch	branch	depth	branch/	depth	branch/	A-A	B-B	A-B	A-B
				length	length		A depth		B depth	distance	distance	distance	distance
global	conserved	ASTRAL	0.975	0.124	0.127	0.902	0.137	0.914	0.135	1.536	1.609	1.973	1.575
global	conserved	CONCAT	1.007	0.126	0.125	0.882	0.143	0.95	0.133	1.533	1.654	1.999	1.576
global	random	ASTRAL	1.787	0.151	0.085	1.406	0.108	1.734	0.087	2.35	3.025	3.3	1.531
global	random	CONCAT	1.809	0.163	0.09	1.391	0.117	1.759	0.093	2.354	3.069	3.321	1.526

Table F.11: Evolutionary proximity between Archaea and Bacteria with 187 extra genomes.

The original 10,575 genomes sampled in March 2017 plus the 187 new genomes sampled in May 2019 which represent previously missing or underrepresented NCBI and GTDB phyla were included in this analysis. The definitions of column names follow Supplementary Table F.4.

Bibliography

[1] Figtree. Available at http://tree.bio.ed.ac.uk/software/figtree/ (Accessed May 2022).

[2] Inmaculada B. Aban and Mark M. Meerschaert. Generalized least-squares estimators for the thickness of heavy tails. *Journal of Statistical Planning and Inference*, 119(2):341–352, 2004.

[3] Andre J. Aberer, Denis Krompass, and Alexandros Stamatakis. Pruning rogue taxa improves phylogenetic accuracy: An efficient algorithm and webservice. *Systematic Biology*, 62(1):162–166, 2013.

[4] Chawin Aiemvaravutigul and Nonthaphat Wongwattanakij. A linear-time algorithm for optimal tree completion. In *2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. IEEE, July 2019.

[5] Orjan Akerborg, Bengt Sennblad, Lars Arvestad, and Jens Lagergren. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proceedings of the National Academy of Sciences*, 106(14):5714–9, apr 2009.

[6] Orjan Akerborg, Bengt Sennblad, and Jens Lagergren. Birth-death prior on phylogeny and speed dating. *BMC Evolutionary Biology*, 8(1):77, 2008.

[7] E S Allman, James H. Degnan, and J A Rhodes. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *J. Math. Biol.*, 62:833–862, 2011.

[8] Elizabeth Allman, James H. Degnan, and John Rhodes. Species tree inference from gene splits by Unrooted STAR methods. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, PP(99):1–7, 2016.

[9] Elizabeth S Allman and John A Rhodes. Phylogenetic invariants for the general Markov model of sequence mutation. *Mathematical biosciences*, 186(2):113–144, 2003.

[10] B Amard and J Bertrand-Sarfati. Microfossils in 2000 ma old cherty stromatolites of the franceville group, gabon. *Precambrian Research*, 81(3-4):197–221, 1997.

[11] Karthik Anantharaman, Christopher T. Brown, Laura A. Hug, Itai Sharon, Cindy J. Castelle, Alexander J. Probst, Brian C. Thomas, Andrea Singh, Michael J. Wilkins, Ulas Karaoz, Eoin L. Brodie, Kenneth H. Williams, Susan S. Hubbard, and Jillian F. Banfield. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature Communications*, 7(1):13219, Oct 2016.

[12] Marti J Anderson. A new method for non-parametric multivariate analysis of variance. *Austral ecology*, 26(1):32–46, 2001.

[13] Dahiana Arcila, Guillermo Ortí, Richard Vari, Jonathan W Armbruster, Melanie L J Stiassny, Kyung D. Ko, Mark H Sabaj, John Lundberg, Liam J Revell, and Ricardo Betancur-R. Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nature Ecology & Evolution*, 1(January):0020, jan 2017.

[14] Stéphane Aris-Brosou. Dating phylogenies with hybrid local molecular clocks. *PLoS One*, 2(9):e879, 2007.

[15] Stéphane Aris-Brosou and Ziheng Yang. Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18s ribosomal RNA phylogeny. *Systematic Biology*, 51(5):703–714, September 2002.

[16] Lars Arvestad, Ann-Charlotte Berglund, Jens Lagergren, and Bengt Sennblad. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. In *Proceedings of the eighth annual international conference on Computational molecular biology - RECOMB '04*, pages 326–335, New York, New York, USA, 2004. ACM Press.

[17] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000.

[18] Mukul S Bansal. Linear-time algorithms for some phylogenetic tree completion problems under robinson-foulds distance. In *RECOMB International conference on Comparative Genomics*, pages 209–226. Springer, 2018.

[19] Susan M Barns, Charles F Delwiche, Jeffrey D Palmer, and Norman R Pace. Perspectives on archaeal diversity, thermophily and monophyly from environmental rrna sequences. *Proceedings of the National Academy of Sciences*, 93(17):9188–9193, 1996.

[20] Daniel Barry and J. A. Hartigan. Statistical Analysis of Hominoid Molecular Evolution. *Statistical Science*, 2(2):191–207, 1987.

[21] Fabia U Battistuzzi and S Blair Hedges. A major clade of prokaryotes with ancient adaptations to life on land. *Molecular biology and evolution*, 26(2):335–343, 2009.

[22] Md. Shamsuzzoha Bayzid and Tandy Warnow. Estimating optimal species trees from incomplete gene trees under deep coalescence. *Journal of computational biology a journal of computational molecular cell biology*, 19(6):591–605, 2012.

[23] Jeremy M. Beaulieu, Brian C. O'Meara, Peter Crane, and Michael J. Donoghue. Heterogeneous Rates of Molecular Evolution and Diversification Could Explain the Triassic Age Estimate for Angiosperms. *Systematic Biology*, 64(5):869–878, 05 2015.

[24] Johannes Bergsten. A review of long-branch attraction. *Cladistics*, 21(2):163–193, 2005.

[25] Holly C Betts, Mark N Puttick, James W Clark, Tom A Williams, Philip CJ Donoghue, and Davide Pisani. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nature ecology & evolution*, 2(10):1556–1562, 2018.

[26] Jacob Bien and Robert Tibshirani. Prototype selection for interpretable classification. *The Annals of Applied Statistics*, 5(4):2403–2424, 2011.

[27] Samuel Blanquart and Nicolas Lartillot. A site- and time-heterogeneous model of amino acid replacement. *Molecular biology and evolution*, 25(5):842–858, 2008.

[28] D. Bogdanowicz and K. Giaro. Matching split distance for unrooted binary phylogenetic trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(1):150–160, Jan 2012.

[29] Damian Bogdanowicz, Krzysztof Giaro, and Borys Wróbel. TreeCmp: Comparison of trees in polynomial time, 2012.

[30] Evan Bolyen, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian C. Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, Eric J. Alm, Manimozhiyan Arumugam, Francesco Asnicar, Yang Bai, Jordan E. Bisanz, Kyle Bittinger, Asker Brejnrod, Colin J. Brislawn, C. Titus Brown, Benjamin J. Callahan, Andrés Mauricio Caraballo-Rodríguez, John Chase, Emily K. Cope, Ricardo Da Silva, Christian Diener, Pieter C. Dorrestein, Gavin M. Douglas, Daniel M. Durall, Claire Duvallet, Christian F. Edwardson, Madeleine Ernst, Mehrbod Estaki, Jennifer Fouquier, Julia M. Gauglitz, Sean M. Gibbons, Deanna L. Gibson, Antonio Gonzalez, Kestrel Gorlick, Jiarong Guo, Benjamin Hillmann, Susan Holmes, Hannes Holste, Curtis Huttenhower, Gavin A. Huttley, Stefan Janssen, Alan K. Jarmusch, Lingjing Jiang, Benjamin D. Kaehler, Kyo Bin Kang, Christopher R. Keefe, Paul Keim, Scott T. Kelley, Dan Knights, Irina Koester, Tomasz Kosciolek, Jorden Kreps, Morgan G. I. Langille, Joslynn Lee, Ruth Ley, Yong-Xin Liu, Erikka Loftfield, Catherine Lozupone, Massoud Maher, Clarisse Marotz, Bryan D. Martin, Daniel McDonald, Lauren J. McIver, Alexey V. Melnik, Jessica L. Metcalf, Sydney C. Morgan, Jamie T. Morton, Ahmad Turan Naimey, Jose A. Navas-Molina, Louis Felix Nothias, Stephanie B. Orchanian, Talima Pearson, Samuel L. Peoples, Daniel Petras, Mary Lai Preuss, Elmar Pruesse, Lasse Buur Rasmussen, Adam Rivers, Michael S. Robeson, Patrick Rosenthal, Nicola Segata, Michael Shaffer, Arron Shiffer, Rashmi Sinha, Se Jin Song, John R. Spear,

Austin D. Swafford, Luke R. Thompson, Pedro J. Torres, Pauline Trinh, Anupriya Tripathi, Peter J. Turnbaugh, Sabah Ul-Hasan, Justin J. J. van der Hooft, Fernando Vargas, Yoshiki Vázquez-Baeza, Emily Vogtmann, Max von Hippel, William Walters, Yunhu Wan, Mingxun Wang, Jonathan Warren, Kyle C. Weber, Charles H. D. Williamson, Amy D. Willis, Zhenjiang Zech Xu, Jesse R. Zaneveld, Yilong Zhang, Qiyun Zhu, Rob Knight, and J. Gregory Caporaso. Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nature Biotechnology*, 37(8):852–857, Aug 2019.

[31] Laura Bonetta. Whole-Genome sequencing breaks the cost barrier. *Cell*, 141(6):917–919, 2010.

[32] B Boussau and M Gouy. Efficient likelihood computations with nonreversible models of evolution. *Systematic biology*, 55(5):756–768, 2006.

[33] Bastien Boussau, GJ J Szöllősi, and Laurent Duret. Genome-scale coestimation of species and gene trees. *Genome Research*, 23(2):323–330, dec 2013.

[34] Robert M. Bowers, Nikos C. Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, T. B. K. Reddy, Frederik Schulz, Jessica Jarett, Adam R. Rivers, Emiley A. Eloe-Fadrosh, Susannah G. Tringe, Natalia N. Ivanova, Alex Copeland, Alicia Clum, Eric D. Becraft, Rex R. Malmstrom, Bruce Birren, Mircea Podar, Peer Bork, George M. Weinstock, George M. Garrity, Jeremy A. Dodsworth, Shibu Yooseph, Granger Sutton, Frank O. Glöckner, Jack A. Gilbert, William C. Nelson, Steven J. Hallam, Sean P. Jungbluth, Thijs J. G. Ettema, Scott Tighe, Konstantinos T. Konstantinidis, Wen-Tso Liu, Brett J. Baker, Thomas Rattei, Jonathan A. Eisen, Brian Hedlund, Katherine D. McMahon, Noah Fierer, Rob Knight, Rob Finn, Guy Cochrane, Ilene Karsch-Mizrachi, Gene W. Tyson, Christian Rinke, Lynn Schriml, Philip Hugenholtz, Pelin Yilmaz, Folker Meyer, Alla Lapidus, Donovan H. Parks, A. Murat Eren, Jillian F. Banfield, Tanja Woyke, and The Genome Standards Consortium. Minimum information about a single amplified genome (misag) and a metagenome-assembled genome (mimag) of bacteria and archaea. *Nature Biotechnology*, 35(8):725–731, Aug 2017.

[35] Laura M Boykin, Laura Salter Kubatko, and Timothy K Lowrey. Comparison of methods for rooting phylogenetic trees: A case study using Orcuttieae (Poaceae: Chloridoideae). *Molecular phylogenetics and evolution*, 54(3):687–700, 2010.

[36] M J Braun, Janice E Clements, and M A Gonda. The visna virus genome: evidence for a hypervariable site in the env gene and sequence homology among lentivirus envelope proteins. *Journal of virology*, 61(12):4046–4054, 1987.

[37] Tom Britton, Cajsa Lisa Anderson, David Jacquet, Samuel Lundqvist, and Kåre Bremer. Estimating Divergence Times in Large Phylogenetic Trees. *Systematic Biology*, 56(5):741–752, 10 2007.

[38] Gerth Stølting Brodal, Rolf Fagerberg, Thomas Mailund, Christian N. S. Pedersen, and Andreas Sand. Efficient Algorithms for Computing the Triplet and Quartet Distance

Between Trees of Arbitrary Degree. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '13, pages 1814–1832, Philadelphia, PA, 1 2013. Society for Industrial and Applied Mathematics.

[39] Lindell Bromham and David Penny. The modern molecular clock. *Nature Reviews Genetics*, 4(3):216–224, 3 2003.

[40] Christopher T Brown, Laura A Hug, Brian C Thomas, Itai Sharon, Cindy J Castelle, Andrea Singh, Michael J Wilkins, Kelly C Wrighton, Kenneth H Williams, and Jillian F Banfield. Unusual biology across a group comprising more than 15% of domain bacteria. *Nature*, 523(7559):208–211, 2015.

[41] James R Brown and W Ford Doolittle. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proceedings of the National Academy of Sciences*, 92(7):2441–2445, 1995.

[42] Richard P Brown and Ziheng Yang. Rate variation and estimation of divergence times using strict and relaxed clocks. *BMC Evolutionary Biology*, 11(1):271, 2011.

[43] Nicholas J Butterfield. Proterozoic photosynthesis–a critical review. *Palaeontology*, 58(6):953–972, 2015.

[44] Sébastien Calvignac-Spencer, Jakob M Schulze, Franziska Zickmann, and Bernhard Y Renard. Clock rooting further demonstrates that guinea 2014 ebov is a member of the zaïre lineage. *PLoS currents*, 6, 2014.

[45] Barbara J Campbell, Annette Summers Engel, Megan L Porter, and Ken Takai. The versatile ε-proteobacteria: key players in sulphidic habitats. *Nature Reviews Microbiology*, 4(6):458–468, 2006.

[46] James H Campbell, Patrick O'Donoghue, Alisha G Campbell, Patrick Schwientek, Alexander Sczyrba, Tanja Woyke, Dieter Söll, and Mircea Podar. Uga is an additional glycine codon in uncultured sr1 bacteria from the human microbiota. *Proceedings of the National Academy of Sciences*, 110(14):5540–5545, 2013.

[47] Johanna Taylor Cannon, Bruno Cossermelli Vellutini, Julian Smith, Fredrik Ronquist, Ulf Jondelius, and Andreas Hejnol. Xenacoelomorpha is the sister group to Nephrozoa. *Nature*, 530(7588):89–93, feb 2016.

[48] Salvador Capella-Gutiérrez, José M Silla-Martínez, and Toni Gabaldón. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973, 2009.

[49] Tanai Cardona. Thinking twice about the evolution of photosynthesis. *Open biology*, 9(3):180246, 2019.

[50] Cindy J. Castelle and Jillian F. Banfield. Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. *Cell*, 172(6):1181–1197, 3 2018.

[51] J. Castresana. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution*, 17(4):540–552, 2000.

[52] Thomas Cavalier-Smith. Rooting the tree of life by transition analyses. *Biology direct*, 1(1):1–83, 2006.

[53] Sarah Christensen, Erin K. Molloy, Pranjal Vachaspati, and Tandy Warnow. OCTAL: Optimal Completion of gene trees in polynomial time. *Algorithms for Molecular Biology*, 13(1):6, 12 2018.

[54] Francesca D Ciccarelli, Tobias Doerks, Christian Von Mering, Christopher J Creevey, Berend Snel, and Peer Bork. Toward automatic reconstruction of a highly resolved tree of life. *science*, 311(5765):1283–1287, 2006.

[55] David Clifford, Noel Cressie, Jacqueline R. England, Stephen H. Roxburgh, and Keryn I. Paul. Correction factors for unbiased, efficient estimation and prediction of biomass from log–log allometric models. *Forest Ecology and Management*, 310:375–381, 2013.

[56] Tom Coenye and Peter Vandamme. A genomic perspective on the relationship between the aquificales and the e-proteobacteria. *Systematic and applied microbiology*, 27(3):313–322, 2004.

[57] Gerard Cornuejols, Marshall L Fisher, and George L Nemhauser. Exceptional paper—location of bank accounts to optimize float: An analytic study of exact and approximate algorithms. *Management science*, 23(8):789–810, 1977.

[58] Cymon J Cox, Peter G Foster, Robert P Hirt, Simon R Harris, and T Martin Embley. The archaebacterial origin of eukaryotes. *Proceedings of the National Academy of Sciences*, 105(51):20356–20361, 2008.

[59] Christopher J Creevey, Tobias Doerks, David A Fitzpatrick, Jeroen Raes, and Peer Bork. Universally distributed single-copy genes indicate a constant rate of horizontal transfer. *PloS one*, 6(8):e22099, 2011.

[60] Sean A Crowe, Lasse N Døssing, Nicolas J Beukes, Michael Bau, Stephanus J Kruger, Robert Frei, and Donald E Canfield. Atmospheric oxygenation three billion years ago. *Nature*, 501(7468):535–538, 2013.

[61] Tal Dagan, Mayo Roettger, David Bryant, and William Martin. Genome networks root the tree of life between prokaryotic domains. *Genome Biology and Evolution*, 2:379–392, 2010.

[62] G Brent Dalrymple. The age of the earth in the twentieth century: a problem (mostly) solved. *Geological Society, London, Special Publications*, 190(1):205–221, 2001.

[63] Aaron E Darling, Guillaume Jospin, Eric Lowe, Frederick A. Matsen, Holly M Bik, and Jonathan A Eisen. PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ*, 2:e243, 1 2014.

[64] Charles Darwin. *The origin of species by means of natural selection*. J. Murray, 1872.

[65] Lawrence a David and Eric J Alm. Rapid evolutionary innovation during an Archaean genetic expansion. *Nature*, 469(7328):93–96, 2011.

[66] Ruth Davidson, Pranjal Vachaspati, Siavash Mirarab, and Tandy Warnow. Phylogenomic species tree estimation in the presence of incomplete lineage sorting and horizontal gene transfer. *BMC genomics*, 16(10):1–12, 2015.

[67] Adrián A Davín, Eric Tannier, Tom A Williams, Bastien Boussau, Vincent Daubin, and Gergely J Szöllősi. Gene transfers can date the tree of life. *Nature ecology & evolution*, 2(5):904–909, 2018.

[68] James H. Degnan and Noah A. Rosenberg. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution*, 24(6):332–340, 6 2009.

[69] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[70] Todd Z DeSantis, Philip Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and Gary L Andersen. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Appl. Environ. Microbiol.*, 72(7):5069–5072, 7 2006.

[71] Sara C Di Rienzi, Itai Sharon, Kelly C Wrighton, Omry Koren, Laura A Hug, Brian C Thomas, Julia K Goodrich, Jordana T Bell, Timothy D Spector, Jillian F Banfield, and Ruth E Ley. The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. *eLife*, 2, 10 2013.

[72] Matthew S Dodd, Dominic Papineau, Tor Grenne, John F Slack, Martin Rittner, Franco Pirajno, Jonathan O'Neil, and Crispin TS Little. Evidence for early life in earth's oldest hydrothermal vent precipitates. *Nature*, 543(7643):60–64, 2017.

[73] CAROL D VON DOHLEN and Nancy A Moran. Molecular data support a rapid radiation of aphids in the cretaceous and multiple origins of host alternation. *Biological Journal of the Linnean Society*, 71(4):689–717, 2000.

[74] Philip C. J. Donoghue and Ziheng Yang. The evolution of methods for establishing evolutionary timescales. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1699):20160020, 7 2016.

[75] Vinson P. Doyle, Randee E. Young, Gavin J.P. Naylor, and Jeremy M. Brown. Can we identify genes with increased phylogenetic reliability? *Systematic Biology*, 64(5):824–837, June 2015.

[76] Alexei J Drummond and Remco R Bouckaert. *Bayesian evolutionary analysis with BEAST*. Cambridge University Press, 2015.

[77] Alexei J. Drummond, Simon Y. W Ho, Matthew J. Phillips, and Andrew Rambaut. Relaxed Phylogenetics and Dating with Confidence. *PLoS Biology*, 4(5):e88, 3 2006.

[78] Alexei J Drummond and Andrew Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, 7:214, 2007.

[79] Alexei J. Drummond and Marc A. Suchard. Bayesian random local clocks, or one rate to rule them all. *BMC Biology*, 8(1):114, 12 2010.

[80] Gytis Dudas, Luiz Max Carvalho, Trevor Bedford, Andrew J. Tatem, Guy Baele, Nuno R. Faria, Daniel J. Park, Jason T. Ladner, Armando Arias, Danny Asogun, Filip Bielejec, Sarah L. Caddy, Matthew Cotten, Jonathan D'Ambrozio, Simon Dellicour, Antonino Di Caro, Joseph W. Diclaro, Sophie Duraffour, Michael J. Elmore, Lawrence S. Fakoli, Ousmane Faye, Merle L. Gilbert, Sahr M. Gevao, Stephen Gire, Adrianne Gladden-Young, Andreas Gnirke, Augustine Goba, Donald S. Grant, Bart L. Haagmans, Julian A. Hiscox, Umaru Jah, Jeffrey R. Kugelman, Di Liu, Jia Lu, Christine M. Malboeuf, Suzanne Mate, David A. Matthews, Christian B. Matranga, Luke W. Meredith, James Qu, Joshua Quick, Suzan D. Pas, My V. T. Phan, Georgios Pollakis, Chantal B. Reusken, Mariano Sanchez-Lockhart, Stephen F. Schaffner, John S. Schieffelin, Rachel S. Sealfon, Etienne Simon-Loriere, Saskia L. Smits, Kilian Stoecker, Lucy Thorne, Ekaete Alice Tobin, Mohamed A. Vandi, Simon J. Watson, Kendra West, Shannon Whitmer, Michael R. Wiley, Sarah M. Winnicki, Shirlee Wohl, Roman Wölfel, Nathan L. Yozwiak, Kristian G. Andersen, Sylvia O. Blyden, Fatorma Bolay, Miles W. Carroll, Bernice Dahn, Boubacar Diallo, Pierre Formenty, Christophe Fraser, George F. Gao, Robert F. Garry, Ian Goodfellow, Stephan Günther, Christian T. Happi, Edward C. Holmes, Brima Kargbo, Sakoba Keïta, Paul Kellam, Marion P. G. Koopmans, Jens H. Kuhn, Nicholas J. Loman, N'Faly Magassouba, Dhamari Naidoo, Stuart T. Nichol, Tolbert Nyenswah, Gustavo Palacios, Oliver G. Pybus, Pardis C. Sabeti, Amadou Sall, Ute Ströher, Isatta Wurie, Marc A. Suchard, Philippe Lemey, and Andrew Rambaut. Virus genomes reveal factors that spread and sustained the ebola epidemic. *Nature*, 544(7650):309–315, 2017.

[81] Barbara Dunn and Gavin Sherlock. Reconstruction of the genome origins and evolution of the hybrid lager yeast Saccharomyces pastorianus. *Genome Research*, 18(10):1610–1623, 8 2008.

[82] Sean R Eddy. A new generation of homology search tools based on probabilistic inference. In *Genome Informatics 2009: Genome Informatics Series Vol. 23*, pages 205–211. World Scientific, 2009.

[83] Scott V Edwards. Is a new and general theory of molecular systematics emerging? *Evolution*, 63(1):1–19, 2009.

[84] Scott V. Edwards, Zhenxiang Xi, Axel Janke, Brant C. Faircloth, John E. McCormack, Travis C. Glenn, Bojian Zhong, Shaoyuan Wu, Emily Moriarty Lemmon, Alan R. Lemmon, Adam D. Leaché, Liang Liu, and Charles C. Davis. Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Molecular phylogenetics and evolution*, 94:447–462, 2016.

[85] Jonathan A. Eisen, Karen E. Nelson, Ian T. Paulsen, John F. Heidelberg, Martin Wu, Robert J. Dodson, Robert Deboy, Michelle L. Gwinn, William C. Nelson, Daniel H. Haft, Erin K. Hickey, Jeremy D. Peterson, A. Scott Durkin, James L. Kolonay, Fan Yang, Ingeborg Holt, Lowell A. Umayam, Tanya Mason, Michael Brenner, Terrance P. Shea, Debbie Parksey, William C. Nierman, Tamara V. Feldblyum, Cheryl L. Hansen, M. Brook Craven, Diana Radune, Jessica Vamathevan, Hoda Khouri, Owen White, Tanja M. Gruber, Karen A. Ketchum, J. Craig Venter, Hervé Tettelin, Donald A. Bryant, and Claire M. Fraser. Chlorobium tepidum tls, a photosynthetic, anaerobic, green-sulfur bacterium. *Proceedings of the National Academy of Sciences*, 99(14):9509–9514, July 2002.

[86] Laura Eme, Anja Spang, Jonathan Lombard, Courtney W Stairs, and Thijs JG Ettema. Archaea and the origin of eukaryotes. *Nature Reviews Microbiology*, 15(12):711–723, 2017.

[87] Peter Erdos, Mike Steel, L Szekely, and Tandy Warnow. A few logs suffice to build (almost) all trees: Part II. *Theoretical Computer Science*, 221(1-2):77–118, 1999.

[88] Jason Evans, Luke Sheneman, and James Foster. Relaxed neighbor joining: a fast distance-based phylogenetic tree construction method. *Journal of molecular evolution*, 62(6):785–792, 2006.

[89] Jason Evans and Jack Sullivan. Approximating model probabilities in bayesian information criterion and decision-theoretic approaches to model selection in phylogenetics. *Molecular biology and evolution*, 28(1):343–349, 2011.

[90] Scott Federhen. The ncbi taxonomy database. *Nucleic acids research*, 40(D1):D136–D143, 2012.

[91] Joseph Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 11 1981.

[92] Joseph Felsenstein. Phylogenies and the Comparative Method. *Am. Nat*, 125(125):3–147, 1985.

[93] Joseph Felsenstein. Phylip (phylogeny inference package) version 3.6. distributed by the author. *http://www. evolution. gs. washington. edu/phylip. html*, 2004.

[94] Joseph Felsenstein and Joseph Felenstein. *Inferring phylogenies*, volume 2. Sinauer associates Sunderland, MA, 2004.

[95] Y Feng, DC Blackburn, D Liang, DM Hillis, DB Wake, DC Cannatella, and P Zhang. Data from: Phylogenomics reveals rapid, simultaneous diversification of three major clades of gondwanan frogs at the cretaceous–paleogene boundary, 2017.

[96] Yan-Jie Feng, David C. Blackburn, Dan Liang, David M. Hillis, David B. Wake, David C. Cannatella, and Peng Zhang. Phylogenomics reveals rapid, simultaneous diversification of three major clades of gondwanan frogs at the cretaceous–paleogene boundary. *Proceedings of the National Academy of Sciences*, 114(29):E5864–E5870, 2017.

[97] Walter M. Fitch. Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology*, 20(4):406–416, 1971.

[98] William Fletcher and Ziheng Yang. INDELible: A flexible simulator of biological sequence evolution. *Molecular biology and evolution*, 26(8):1879–1888, 2009.

[99] Félix Forest. Calibrating the Tree of Life: fossils, molecules and evolutionary timescales. *Annals of Botany*, 104(5):789–794, 10 2009.

[100] Patrick Forterre and Herve Philippe. Where is the root of the universal tree of life? *Bioessays*, 21(10):871–879, 1999.

[101] Mathieu Fourment and Edward C Holmes. Novel non-parametric models to estimate evolutionary rates and divergence times from heterochronous sequence data. *BMC evolutionary biology*, 14(1):1–12, 2014.

[102] Gregory P Fournier and J Peter Gogarten. Rooting the ribosomal tree of life. *Molecular biology and evolution*, 27(8):1792–1801, 2010.

[103] Nicolas Galtier and Vincent Daubin. Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1512):4023–4029, 2008.

[104] J Esteban Gamez, François Modave, and Olga Kosheleva. Selecting the most representative sample is np-hard: Need for expert (fuzzy) knowledge. In *2008 IEEE International Conference on Fuzzy Systems (IEEE World Congress on Computational Intelligence)*, pages 1069–1074. IEEE, 2008.

[105] O Gascuel. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14(7):685–695, 07 1997.

[106] John Gatesy and Mark S Springer. Phylogenetic Analysis at Deep Timescales: Unreliable Gene Trees, Bypassed Hidden Support, and the Coalescence/Concatalescence Conundrum. *Molecular phylogenetics and evolution*, 80:231–266, 2014.

[107] Henry Gee. Evolution: ending incongruence. *Nature*, 425(6960):782, 2003.

[108] Tanja Gernhard. The conditioned reconstructed process. *Journal of Theoretical Biology*, 253(4):769–778, 2008.

[109] Timothy M. Gibson, Patrick M. Shih, Vivien M. Cumming, Woodward W. Fischer, Peter W. Crockford, Malcolm S.W. Hodgskiss, Sarah Wörndle, Robert A. Creaser, Robert H. Rainbird, Thomas M. Skulski, and Galen P. Halverson. Precise age of bangiomorpha pubescens dates the origin of eukaryotic photosynthesis. *Geology*, 46(2):135–138, December 2017.

[110] M. T. P. Gilbert, A. Rambaut, G. Wlasiuk, T. J. Spira, A. E. Pitchenik, and M. Worobey. The emergence of HIV/AIDS in the Americas and beyond. *Proceedings of the National Academy of Sciences*, 104(47):18566–18570, 11 2007.

[111] J. Peter Gogarten, W. Ford Doolittle, and Jeffrey G. Lawrence. Prokaryotic Evolution in Light of Gene Transfer. *Molecular Biology and Evolution*, 19(12):2226–2238, 12 2002.

[112] Pablo A. Goloboff and Claudia A. Szumik. Identifying unstable taxa: Efficient implementation of triplet-based measures of stability, and comparison with Phyutility and RogueNaRok. *Molecular Phylogenetics and Evolution*, 88:93–104, 2015.

[113] Stjepko Golubic and HJ Hofmann. Comparison of holocene and mid-precambrian entophysalidaceae (cyanophyta) in stromatolitic algal mats: cell division and degradation. *Journal of Paleontology*, pages 1074–1082, 1976.

[114] Stjepko Golubic, Vladimir N Sergeev, and Andrew H Knoll. Mesoproterozoic archaeoellipsoides: akinetes of heterocystous cyanobacteria. *Lethaia*, 28(4):285–298, 1995.

[115] Richard Gouy, Denis Baurain, and Hervé Philippe. Rooting the tree of life: the phylogenetic jury is still out. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678):20140329, sep 2015.

[116] Sean W Graham, Richard G Olmstead, and Spencer C H Barrett. Rooting phylogenetic trees with distant outgroups: a case study from the commelinoid monocots. *Molecular biology and evolution*, 19(10):1769–1781, 2002.

[117] Tanja M Gruber and Donald A Bryant. Characterization of the group 1 and group 2 sigma factors of the green sulfur bacterium chlorobium tepidum and the green non-sulfur bacterium chloroflexus aurantiacus. *Archives of microbiology*, 170(4):285–296, 1998.

[118] Stéphane Guindon, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic biology*, 59(3):307–321, may 2010.

[119] Stéphane Guindon. Bayesian Estimation of Divergence Times from Large Sequence Alignments. *Molecular Biology and Evolution*, 27(8):1768–1781, 03 2010.

[120] Radhey S Gupta and Beile Gao. Phylogenomic analyses of clostridia and identification of novel protein signatures that are specific to the genus clostridiumsensu stricto (cluster i). *International journal of systematic and evolutionary microbiology*, 59(2):285–294, 2009.

[121] Lionel Guy and Thijs JG Ettema. The archaeal 'tack'superphylum and the origin of eukaryotes. *Trends in microbiology*, 19(12):580–587, 2011.

[122] Daniel H Haft, Michael DiCuccio, Azat Badretdin, Vyacheslav Brover, Vyacheslav Chetvernin, Kathleen O'Neill, Wenjun Li, Farideh Chitsaz, Myra K Derbyshire, Noreen R Gonzales, Marc Gwadz, Fu Lu, Gabriele H Marchler, James S Song, Narmada Thanki, Roxanne A Yamashita, Chanjuan Zheng, Françoise Thibaud-Nissen, Lewis Y Geer, Aron Marchler-Bauer, and Kim D Pruitt. RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Research*, 46(D1):D851–D860, November 2017.

[123] V. Hampl, L. Hug, J. W. Leigh, J. B. Dacks, B. F. Lang, A. G. B. Simpson, and A. J. Roger. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *Proceedings of the National Academy of Sciences*, 106(10):3859–3864, 2009.

[124] J Kirk Harris, Scott T Kelley, George B Spiegelman, and Norman R Pace. The genetic core of the universal ancestor. *Genome research*, 13(3):407–412, 2003.

[125] Tracy A Heath. A hierarchical Bayesian model for calibrating estimates of species divergence times. *Systematic biology*, 61(5):793–809, 10 2012.

[126] Tracy A Heath, John P Huelsenbeck, and Tanja Stadler. The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences*, 111(29):E2957–E2966, 2014.

[127] J. Hedge, S. J. Lycett, and A. Rambaut. Real-time characterization of the molecular epidemiology of an influenza pandemic. *Biology Letters*, 9(5):20130331, 10 2013.

[128] S Blair Hedges and Sudhir Kumar. *The timetree of life*. OUP Oxford, 2009.

[129] Pablo N. Hess and Claudia A. De Moraes Russo. An empirical test of the midpoint rooting method. *Biological Journal of the Linnean Society*, 92(4):669–674, 2007.

[130] David M Hillis, C Moritz, and B K Mable. *Molecular Systematics*, volume 2nd. Sinauer Associates, Sunderland, MA, 1996.

[131] David M Hillis, David D Pollock, Jimmy A McGuire, and Derrick Joel Zwickl. Is sparse taxon sampling a problem for phylogenetic inference? *Systematic biology*, 52(1):124–126, 2003.

[132] Cody E. Hinchliff, Stephen A. Smith, James F. Allman, J. Gordon Burleigh, Ruchi Chaudhary, Lyndon M. Coghill, Keith A. Crandall, Jiabin Deng, Bryan T. Drew, Romina Gazis, Karl Gude, David S. Hibbett, Laura A. Katz, H. Dail Laughinghouse, Emily Jane McTavish,

Peter E. Midford, Christopher L. Owen, Richard H. Ree, Jonathan A. Rees, Douglas E. Soltis, Tiffani Williams, and Karen A. Cranston. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences*, 112(41):12764–12769, September 2015.

[133] Simon Y. W. Ho, Sebastián Duchêne, and David Duchêne. Simulating and detecting autocorrelation of molecular evolutionary rates among lineages. *Molecular Ecology Resources*, 15(4):688–696, 2015.

[134] Simon Y W Ho and Matthew J. Phillips. Accounting for Calibration Uncertainty in Phylogenetic Estimation of Evolutionary Divergence Times. *Systematic Biology*, 58(3):367–380, 7 2009.

[135] Simon Y.W. Ho. The changing face of the molecular evolutionary clock. *Trends in Ecology & Evolution*, 29(9):496–503, 2014.

[136] Diep Thi Hoang, Olga Chernomor, Arndt Von Haeseler, Bui Quang Minh, and Le Sy Vinh. Ufboot2: improving the ultrafast bootstrap approximation. *Molecular biology and evolution*, 35(2):518–522, 2018.

[137] B R Holland, D Penny, and M D Hendy. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock–a simulation study. *Systematic biology*, 52(2):229–238, 2003.

[138] Robert J Horodyski and J Allan Donaldson. Microfossils from the middle proterozoic dismal lakes groups, arctic canada. *Precambrian Research*, 11(2):125–159, 1980.

[139] Peter A Hosner, Edward L Braun, and Rebecca T Kimball. Land connectivity changes and global cooling shaped the colonization history and diversification of New World quail (Aves : Galliformes : Odontophoridae). *Journal of Biogeography*, 42(10):1–13, 10 2015.

[140] Peter A Hosner, Brant C Faircloth, Travis C. Glenn, Edward L Braun, and Rebecca T Kimball. Avoiding Missing Data Biases in Phylogenomic Inference: An Empirical Study in the Landfowl (Aves: Galliformes). *Molecular Biology and Evolution*, 33(4):1110–1125, 4 2016.

[141] Rasmus Hovmöller, L. Lacey Knowles, and Laura S. Kubatko. Effects of missing data on species tree estimation under the coalescent. *Molecular Phylogenetics and Evolution*, 69(3):1057–1062, December 2013.

[142] John P Huelsenbeck, Jonathan P Bollback, and Amy M Levine. Inferring the root of a phylogenetic tree. *Systematic biology*, 51(1):32–43, 2002.

[143] John P. Huelsenbeck, Bret Larget, and David Swofford. A compound poisson process for relaxing the molecular clock. *Genetics*, 154(4):1879–92, 4 2000.

[144] Laura A. Hug, Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Cristina N. Butterfield, Alex W. Hernsdorf, Yuki Amano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A. Relman, Kari M. Finstad, Ronald Amundson, Brian C. Thomas, and Jillian F. Banfield. A new view of the tree of life. *Nature Microbiology*, 1(5), April 2016.

[145] Philip Hugenholtz and Thomas Huber. Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *International Journal of Systematic and Evolutionary Microbiology*, 53(1):289–293, 2003.

[146] Daniel H Huson and Celine Scornavacca. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Systematic biology*, 61(6):1061–1067, 2012.

[147] Doug Hyatt, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11(1):1–11, 2010.

[148] Ravi Jain, Maria C Rivera, and James A Lake. Horizontal gene transfer among genomes: the complexity hypothesis. *Proceedings of the National Academy of Sciences*, 96(7):3801–3806, 1999.

[149] Erich D Jarvis, Siavash Mirarab, Andre J Aberer, Bo Li, Peter Houde, Cai Li, Simon Y W Ho, Brant C. Faircloth, Benoit Nabholz, Jason T Howard, Alexander Suh, Claudia C Weber, Rute R da Fonseca, Jianwen Li, Fang Zhang, Hui Li, Long Zhou, Nitish Narula, Liang Liu, Ganeshkumar Ganapathy, Bastien Boussau, Md. Shamsuzzoha Bayzid, Volodymyr Zavidovych, Sankar Subramanian, Toni Gabaldón, Salvador Capella-Gutiérrez, Jaime Huerta-Cepas, Bhanu Rekepalli, Kasper Munch, Mikkel H. Schierup, Bent Lindow, Wesley C Warren, David Ray, Richard E Green, Michael W Bruford, Xiangjiang Zhan, Andrew Dixon, Shengbin Li, Ning Li, Yinhua Huang, Elizabeth P Derryberry, Mads Frost Bertelsen, Frederick H Sheldon, Robb T. Brumfield, Claudio V Mello, Peter V Lovell, Morgan Wirthlin, Maria Paula Cruz Schneider, Francisco Prosdocimi, José Alfredo Samaniego, Amhed Missael Vargas Velazquez, Alonzo Alfaro-Núñez, Paula F Campos, Bent Petersen, Thomas Sicheritz-Ponten, An Pas, Tom Bailey, Paul Scofield, Michael Bunce, David M Lambert, Qi Zhou, Polina Perelman, Amy C. Driskell, Beth Shapiro, Zijun Xiong, Yongli Zeng, Shiping Liu, Zhenyu Li, Binghang Liu, Kui Wu, Jin Xiao, Xiong Yinqi, Qiuemei Zheng, Yong Zhang, Huanming Yang, Jian Wang, Linnea Smeds, Frank E Rheindt, Michael J Braun, Jon Fjeldsa, Ludovic Orlando, F Keith Barker, Knud Andreas Jønsson, Warren Johnson, Klaus-Peter Koepfli, Stephen O'Brien, David Haussler, Oliver A Ryder, Carsten Rahbek, Eske Willerslev, Gary R Graves, Travis C. Glenn, John E McCormack, Dave W Burt, Hans Ellegren, Per Alström, Scott V Edwards, Alexandros Stamatakis, David P Mindell, Joel Cracraft, Edward L Braun, Tandy Warnow, Wang Jun, M Thomas P Gilbert, and Guojie Zhang. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331, dec 2014.

[150] Vivek Jayaswal, Lars S Jermiin, and John Robinson. Estimation of phylogeny using a general Markov model. *Evolutionary bioinformatics online*, 1:62–80, 2005.

[151] Olivier Jeffroy, Henner Brinkmann, Frédéric Delsuc, and Hervé Philippe. Phylogenomics: the beginning of incongruence? *Trends in Genetics*, 22(4):225–231, 2006.

[152] Yueyu Jiang, Metin Balaban, Qiyun Zhu, and Siavash Mirarab. DEPP: Deep Learning Enables Extending Species Trees using Single Genes. *bioRxiv (abstract in RECOMB 2021)*, page 2021.01.22.427808, 1 2021.

[153] Jens Johansen and Morten Kragelund Holt. Computing triplet and quartet distances. Master's thesis, Department of Computer Science, Aarhus University, 2013.

[154] Kevin P. Johnson, Christopher H. Dietrich, Frank Friedrich, Rolf G. Beutel, Benjamin Wipfler, Ralph S. Peters, Julie M. Allen, Malte Petersen, Alexander Donath, Kimberly K. O. Walden, Alexey M. Kozlov, Lars Podsiadlowski, Christoph Mayer, Karen Meusemann, Alexandros Vasilikopoulos, Robert M. Waterhouse, Stephen L. Cameron, Christiane Weirauch, Daniel R. Swanson, Diana M. Percy, Nate B. Hardy, Irene Terry, Shanlin Liu, Xin Zhou, Bernhard Misof, Hugh M. Robertson, and Kazunori Yoshizawa. Phylogenomics and the evolution of hemipteroid insects. *Proceedings of the National Academy of Sciences*, 115(50):12775–12780, November 2018.

[155] Eric Jones, Travis Oliphant, and Pearu Peterson. Scipy: Open source scientific tools for python. 2001.

[156] Gregory Jordan and Nick Goldman. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Molecular biology and evolution*, 29(4):1125–1139, 2011.

[157] T H Jukes and C R Cantor. Evolution of protein molecules. In *In Mammalian protein metabolism, Vol. III (1969), pp. 21-132*, volume III, pages 21–132. 1969.

[158] Thomas H Jukes. A change in the genetic code inmycoplasma capricolum. *Journal of Molecular Evolution*, 22(4):361–362, 1985.

[159] Estelle Jumas-Bilak, Laurent Roudiere, and Helene Marchandin. Description of 'synergistetes' phyl. nov. and emended description of the phylum 'deferribacteres' and of the family syntrophomonadaceae, phylum 'firmicutes'. *International journal of systematic and evolutionary microbiology*, 59(5):1028–1035, 2009.

[160] Nahomi Kaiwa, Takahiro Hosokawa, Naruo Nikoh, Masahiko Tanahashi, Minoru Moriyama, Xian-Ying Meng, Taro Maeda, Katsushi Yamaguchi, Shuji Shigenobu, Motomi Ito, and Takema Fukatsu. Symbiont-supplemented maternal investment underpinning host's ecological adaptation. *Current Biology*, 24(20):2465–2470, October 2014.

[161] Subha Kalyaanamoorthy, Bui Quang Minh, Thomas KF Wong, Arndt Von Haeseler, and Lars S Jermiin. Modelfinder: fast model selection for accurate phylogenetic estimates. *Nature methods*, 14(6):587–589, 2017.

[162] Kazutaka Katoh, Kei-ichi Kuma, Hiroyuki Toh, and Takashi Miyata. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucl Acids Res*, 33(2):511–518, 2005.

[163] Laura A. Katz, Jessica R. Grant, Laura Wegener Parfrey, and J. Gordon Burleigh. Turning the crown upside down: Gene tree parsimony roots the eukaryotic tree of life. *Systematic biology*, 61(4):653–660, 2012.

[164] Oliver N. Keene. The log transformation is special. *Statistics in Medicine*, 14(8):811–819, 1995.

[165] Frank Keul, Martin Hess, Michael Goesele, and Kay Hamacher. Pfasum: a substitution matrix from pfam structural alignments. *BMC bioinformatics*, 18(1):1–14, 2017.

[166] Hirohisa Kishino, Jeffrey L. Thorne, and William J. Bruno. Performance of a Divergence Time Estimation Method under a Probabilistic Model of Rate Evolution. *Molecular Biology and Evolution*, 18(3):352–361, 3 2001.

[167] Hans-Peter Klenk, Thomas-Dirk Meier, Peter Durovic, Volker Schwass, Friedrich Lottspeich, Patrick P Dennis, and Wolfram Zillig. Rna polymerase of aquifex pyrophilus: implications for the evolution of the bacterial rpobc operon and extremely thermophilic bacteria. *Journal of molecular evolution*, 48(5):528–541, 1999.

[168] Ullasa Kodandaramaiah. Tectonic calibrations in molecular dating. *Current Zoology*, 57(1):116–124, 2 2011.

[169] Dirk Krüger and Andrea Gargas. New measures of topological stability in phylogenetic trees - Taking taxon composition into account. *Bioinformation*, 1(8):327–30, 2006.

[170] Sudhir Kumar. Molecular clocks: four decades of evolution. *Nature Reviews Genetics*, 6(8):654–662, 8 2005.

[171] Sudhir Kumar and S. Blair Hedges. Advances in Time Estimation Methods for Molecular Data. *Molecular Biology and Evolution*, 33(4):863–869, 02 2016.

[172] Manuel Lafond and Celine Scornavacca. On the Weighted Quartet Consensus problem. *Theoretical Computer Science*, 769:1–17, 5 2019.

[173] James A Lake, Eric Henderson, Melanie Oakes, and Michael W Clark. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proceedings of the National Academy of Sciences*, 81(12):3786–3790, 1984.

[174] James A Lake, Ryan G Skophammer, Craig W Herbold, and Jacqueline A Servin. Genome beginnings: rooting the tree of life. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1527):2177–2185, 2009.

[175] Marucha Lalee, Jorge Nocedal, and Todd Plantenga. On the implementation of an algorithm for large-scale equality constrained optimization. *SIAM Journal on Optimization*, 8(3):682–706, 1998.

[176] Miriam L Land, Doug Hyatt, Se-Ran Jun, Guruprasad H Kora, Loren J Hauser, Oksana Lukjancenko, and David W Ussery. Quality scores for 32,000 genomes. *Standards in genomic sciences*, 9(1):1–10, 2014.

[177] Charles H. Langley and Walter M. Fitch. An examination of the constancy of the rate of molecular evolution. *Journal of Molecular Evolution*, 3(3):161–177, Sep 1974.

[178] Nicolas Lartillot, Henner Brinkmann, and Hervé Philippe. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology*, 7(Suppl 1):S4, 2007.

[179] Nicolas Lartillot and Hervé Philippe. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–1109, 2004.

[180] Simon Laurin-Lemay, Henner Brinkmann, and Hervé Philippe. Origin of land plants revisited in the light of sequence contamination and missing data. *Current Biology*, 2012.

[181] Si Quang Le and Olivier Gascuel. An improved general amino acid replacement matrix. *Molecular biology and evolution*, 25(7):1307–1320, 2008.

[182] Adam D Leaché and Bruce Rannala. The accuracy of species tree estimation under simulation: A comparison of methods. *Systematic biology*, 60(2):126–137, mar 2011.

[183] Vincent Lefort, Richard Desper, and Olivier Gascuel. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. *Molecular Biology and Evolution*, 32(10):2798–2800, 06 2015.

[184] Philippe Lemey, Marc Suchard, and Andrew Rambaut. Reconstructing the initial global spread of a human influenza pandemic: A bayesian spatial-temporal model for the global spread of h1n1pdm. *PLoS currents*, 1:RRN1031–RRN1031, Sep 2009.

[185] Thomas Lepage, David Bryant, Hervé Philippe, and Nicolas Lartillot. A General Comparison of Relaxed Molecular Clock Models. *Molecular Biology and Evolution*, 24(12):2669–2680, 6 2007.

[186] Ivica Letunic and Peer Bork. Interactive tree of life (itol) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic acids research*, 44(W1):W242–W245, 2016.

[187] Chenhong Li, Kerri A Matthes-Rosana, Michael Garcia, and Gavin J P Naylor. Phylogenetics of Chondrichthyes and the problem of rooting phylogenies with distant outgroups. *Molecular phylogenetics and evolution*, 63(2):365–373, may 2012.

[188] C Randal Linder and Tandy Warnow. An overview of phylogeny reconstruction. 2001.

[189] Susan Little, Sergei L Kosakovsky Pond, Christy M. Anderson, Jason A. Young, Joel O. Wertheim, Sanjay R. Mehta, Susanne J May, and Davey M. Smith. Using HIV networks to inform real time prevention interventions. *PLoS ONE*, 9(6), 2014.

[190] Bo Liu, Theodore Gibbons, Mohammad Ghodsi, and Mihai Pop. MetaPhyler: Taxonomic profiling for metagenomic sequences. In *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on*, page 95–100. IEEE, 2011.

[191] Liang Liu and Lili Yu. Estimating species trees from unrooted gene trees. *Systematic biology*, 60(5):661–667, 10 2011.

[192] Liang Liu, Lili Yu, and Scott V Edwards. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, 10(1):302, 2010.

[193] Liang Liu, Lili Yu, Dennis K Pearl, and Scott V Edwards. Estimating species phylogenies using coalescence times among sequences. *Systematic biology*, 58(5):468–477, oct 2009.

[194] Genming Luo, Shuhei Ono, Nicolas J Beukes, David T Wang, Shucheng Xie, and Roger E Summons. Rapid oxygenation of earth's atmosphere 2.33 billion years ago. *Science Advances*, 2(5):e1600134, 2016.

[195] Timothy W Lyons, Christopher T Reinhard, and Noah J Planavsky. The rise of oxygen in earth's early ocean and atmosphere. *Nature*, 506(7488):307–315, 2014.

[196] Wayne P. Maddison. Gene Trees in Species Trees. *Systematic biology*, 46(3):523–536, sep 1997.

[197] Wayne P. Maddison, Michael J. Donoghue, and David R. Maddison. Outgroup analysis and parsimony. *Systematic biology*, 33(1):83–103, 1984.

[198] C Magnabosco, Kelsey Reed Moore, Joanna Michelle Wolfe, and Gregory P Fournier. Dating phototrophic microbial lineages with reticulate gene histories. *Geobiology*, 16(2):179–189, 2018.

[199] Uyen Mai and Siavash Mirarab. TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics*, 19(S5):272, 5 2018.

[200] Uyen Mai and Siavash Mirarab. Log Transformation Improves Dating of Phylogenies. *Molecular Biology and Evolution*, 38(3):1151–1167, 09 2020.

[201] Uyen Mai, Erfan Sayyari, and Siavash Mirarab. Minimum variance rooting of phylogenetic trees and implications for species tree reconstruction. *PLOS ONE*, 12(8):1–19, 08 2017.

[202] Diego Mallo, Leonardo de Oliveira Martins, and David Posada. SimPhy: Phylogenomic Simulation of Gene, Locus and Species Trees. *Systematic biology*, 65(2):syv082–, jun 2016.

[203] Julie Marin, Fabia U Battistuzzi, Anais C Brown, and S Blair Hedges. The timetree of prokaryotes: new insights into their evolution and speciation. *Molecular biology and evolution*, 34(2):437–446, 2016.

[204] Joran Martijn, Julian Vosseberg, Lionel Guy, Pierre Offre, and Thijs JG Ettema. Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature*, 557(7703):101–105, 2018.

[205] DésiréL Massart, Frank Plastria, and Leonard Kaufman. Non-hierarchical clustering with masloc. *Pattern Recognition*, 16(5):507–516, 1983.

[206] Sarah Mathews and Michael J. Donoghue. The Root of Angiosperm Phylogeny Inferred from Duplicate Phytochrome Genes. *Science*, 286(1999):947–950, 1999.

[207] Frederick A Matsen, Robin B Kodner, and E Virginia Armbrust. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics*, 11(1):538, jan 2010.

[208] John P McCutcheon, Bradon R McDonald, and Nancy A Moran. Origin of an alternative genetic code in the extremely small and gc–rich genome of a bacterial symbiont. *PLoS genetics*, 5(7):e1000565, 2009.

[209] Daniel McDonald, Morgan N Price, Julia Goodrich, Eric P Nawrocki, Todd Z DeSantis, Alexander Probst, Gary L Andersen, Rob Knight, and Philip Hugenholtz. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME journal*, 6(3):610–618, 3 2012.

[210] Bryan S Mclean, Kayce C Bell, Julie M Allen, Kristofer M Helgen, and Joseph A Cook. Impacts of inference method and data set filtering on phylogenomic resolution in a rapid radiation of ground squirrels (xerinae: Marmotini). *Systematic Biology*, 68(2):298–316, September 2018.

[211] Paul M. Meaney, Qianqian Fang, Tonny Rubaek, Eugene Demidenko, and Keith D. Paulsen. Log transformation benefits parameter estimation in microwave tomographic imaging. *Medical Physics*, 34(6Part1):2014–2023, May 2007.

[212] Kelly A. Meiklejohn, Brant C. Faircloth, Travis C. Glenn, Rebecca T. Kimball, and Edward L. Braun. Analysis of a Rapid Evolutionary Radiation Using Ultraconserved Elements: Evidence for a Bias in Some Multispecies Coalescent Methods. *Systematic biology*, 65(4):612–627, jul 2016.

[213] Michael L Metzker. Sequencing technologies—the next generation. *Nature reviews genetics*, 11(1):31–46, 2010.

[214] S. Mirarab, M.S. Bayzid, B. Boussau, and T. Warnow. Response to Comment on ”Statistical binning enables an accurate coalescent-based estimation of the avian tree”. *Science*, 350(6257), 2015.

[215] Siavash Mirarab. Novel scalable approaches for multiple sequence alignment and phylogenomic reconstruction, 2015.

[216] Siavash Mirarab, Shamsuzzoha Md Bayzid, Bastien Boussau, and Tandy Warnow. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science*, 346(6215):1250463–1250463, dec 2014.

[217] Siavash Mirarab, Nam Nguyen, Sheng Guo, Li-San Wang, Junhyong Kim, and Tandy Warnow. PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences. *Journal of Computational Biology*, 22(05):377–386, 5 2015.

[218] Siavash Mirarab, Rezwana Reaz, Md. Shamsuzzoha Bayzid, Théo Zimmermann, M. S. Swenson, and Tandy Warnow. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548, sep 2014.

[219] Siavash Mirarab and Tandy Warnow. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, 31(12):i44–i52, jun 2015.

[220] Bernhard Misof, Shanlin Liu, Karen Meusemann, Ralph S Peters, Alexander Donath, Christoph Mayer, Paul B. Frandsen, Jessica Ware, Tomáš Flouri, Rolf G Beutel, Oliver Niehuis, Malte Petersen, F. Izquierdo-Carrasco, T. Wappler, J. Rust, A. J. Aberer, U. Aspock, H. Aspock, D. Bartel, A. Blanke, S. Berger, A. Bohm, T. R. Buckley, B. Calcott, J. Chen, F. Friedrich, M. Fukui, M. Fujita, C. Greve, P. Grobe, S. Gu, Y. Huang, L. S. Jermiin, A. Y. Kawahara, L. Krogmann, M. Kubiak, R. Lanfear, H. Letsch, Y. Li, Z. Li, J. Li, H. Lu, R. Machida, Y. Mashimo, P. Kapli, D. D. McKenna, G. Meng, Y. Nakagaki, J. L. Navarrete-Heredia, M. Ott, Y. Ou, G. Pass, L. Podsiadlowski, H. Pohl, B. M. von Reumont, K. Schutte, K. Sekiya, S. Shimizu, A. Slipinski, A. Stamatakis, W. Song, X. Su, N. U. Szucsich, M. Tan, X. Tan, M. Tang, J. Tang, G. Timelthaler, S. Tomizuka, M. Trautwein, X. Tong, T. Uchifune, M. G. Walzl, B. M. Wiegmann, J. Wilbrandt, B. Wipfler, T. K. F. Wong, Q. Wu, G. Wu, Y. Xie, S. Yang, Q. Yang, D. K. Yeates, K. Yoshizawa, Q. Zhang, R. Zhang, W. Zhang, Y. Zhang, J. Zhao, C. Zhou, L. Zhou, T. Ziesmann, S. Zou, Y. Li, X. Xu, Y. Zhang, H. Yang, J. Wang, J. Wang, K. M. Kjer, and X. Zhou. Phylogenomics resolves the timing and pattern of insect evolution. *Science*, 346(6210):763–767, 11 2014.

[221] Aleksandra M. Mloszewska, Devon B. Cole, Noah J. Planavsky, Andreas Kappler, Denise S. Whitford, George W. Owttrim, and Kurt. O Konhauser. UV radiation limited the expansion of cyanobacteria in early marine photic environments. *Nature Communications*, 9(1), August 2018.

[222] Erin K. Molloy and Tandy Warnow. To Include or Not to Include: The Impact of Gene Filtering on Species Tree Estimation Methods. *Systematic Biology*, 67(2):285–303, 3 2018.

[223] David Moreira, Herve Le Guyader, and Herve Philippe. The origin of red algae and the evolution of chloroplasts. *Nature*, 405(6782):69–72, 2000.

[224] Elchanan Mossel and Sebastien Roch. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(1):166–171, jan 2010.

[225] Supratim Mukherjee, Rekha Seshadri, Neha J Varghese, Emiley A Eloe-Fadrosh, Jan P Meier-Kolthoff, Markus Göker, R Cameron Coates, Michalis Hadjithomas, Georgios A Pavlopoulos, David Paez-Espino, Yasuo Yoshikuni, Axel Visel, William B Whitman, George M Garrity, Jonathan A Eisen, Philip Hugenholtz, Amrita Pati, Natalia N Ivanova, Tanja Woyke, Hans-Peter Klenk, and Nikos C Kyrpides. 1, 003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nature Biotechnology*, 35(7):676–683, June 2017.

[226] Ahmed Ragab Nabhan and Indra Neil Sarkar. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Briefings in bioinformatics*, 13(1):122–134, 2012.

[227] D Nguyen Nam-phuong, Siavash Mirarab, Keerthana Kumar, and Tandy Warnow. Ultra-large alignments using phylogeny-aware profiles. *Genome biology*, 16(1):1–15, 2015.

[228] Stephen G Nash. A survey of truncated-newton methods. *Journal of computational and applied mathematics*, 124(1-2):45–59, 2000.

[229] S. Nee, R. M. May, and P. H. Harvey. The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1309):305–311, 5 1994.

[230] Serita M Nelesen, Kevin Liu, Li-San Wang, C. Randal Linder, and Tandy Warnow. DACTAL: divide-and-conquer trees (almost) without alignments. *Bioinformatics*, 28(12):i274—-i282, 2012.

[231] Lam Tung Nguyen, Heiko A. Schmidt, Arndt Von Haeseler, and Bui Quang Minh. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1), 2015.

[232] Nam Nguyen, Siavash Mirarab, Bo Liu, Mihai Pop, and Tandy Warnow. TIPP: Taxonomic Identification and Phylogenetic Profiling. *Bioinformatics*, 30(24):3548–3555, 12 2014.

[233] H Bjørn Nielsen, , Mathieu Almeida, Agnieszka Sierakowska Juncker, Simon Rasmussen, Junhua Li, Shinichi Sunagawa, Damian R Plichta, Laurent Gautier, Anders G Pedersen, Emmanuelle Le Chatelier, Eric Pelletier, Ida Bonde, Trine Nielsen, Chaysavanh Manichanh,

Manimozhiyan Arumugam, Jean-Michel Batto, Marcelo B Quintanilha dos Santos, Nikolaj Blom, Natalia Borruel, Kristoffer S Burgdorf, Fouad Boumezbeur, Francesc Casellas, Joël Doré, Piotr Dworzynski, Francisco Guarner, Torben Hansen, Falk Hildebrand, Rolf S Kaas, Sean Kennedy, Karsten Kristiansen, Jens Roat Kultima, Pierre Léonard, Florence Levenez, Ole Lund, Bouziane Moumen, Denis Le Paslier, Nicolas Pons, Oluf Pedersen, Edi Prifti, Junjie Qin, Jeroen Raes, Søren Sørensen, Julien Tap, Sebastian Tims, David W Ussery, Takuji Yamada, Pierre Renault, Thomas Sicheritz-Ponten, Peer Bork, Jun Wang, Søren Brunak, and S Dusko Ehrlich. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology*, 32(8):822–828, July 2014.

[234] EG Nisbet, NV Grassineau, CJ Howe, PI Abell, M Regelous, and RER Nisbet. The age of rubisco: the evolution of oxygenic photosynthesis. *Geobiology*, 5(4):311–335, 2007.

[235] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, NY, USA, 2e edition, 2006.

[236] Imen Nouioui, Lorena Carro, Marina García-López, Jan P Meier-Kolthoff, Tanja Woyke, Nikos C Kyrpides, Rüdiger Pukall, Hans-Peter Klenk, Michael Goodfellow, and Markus Göker. Genome-based taxonomic classification of the phylum actinobacteria. *Frontiers in microbiology*, page 2007, 2018.

[237] Michael Nute, Jed Chou, Erin K. Molloy, and Tandy Warnow. The performance of coalescent-based species tree estimation methods under models of missing data. *BMC Genomics*, 19(S5):286, 5 2018.

[238] Jesús AG Ochoa de Alda, Rocío Esteban, María Luz Diago, and Jean Houmard. The plastid ancestor originated among one of the major cyanobacterial lineages. *Nature communications*, 5(1):1–10, 2014.

[239] Maureen A O'Malley and Eugene V Koonin. How stands the tree of life a century and a half after the origin? *Biology Direct*, 6(1):1–21, 2011.

[240] Brian D Ondov, Todd J Treangen, Páll Melsted, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, and Adam M Phillippy. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 17(1):132, 12 2016.

[241] One Thousand Plant Transcriptomes OneKP Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, 574(7780):679–685, 2019.

[242] Frantz Ossa Ossa, Axel Hofmann, Jorge E Spangenberg, Simon W Poulton, Eva E Stüeken, Ronny Schoenberg, Benjamin Eickmann, Martin Wille, Mike Butler, and Andrey Bekker. Limited oxygen production in the mesoarchean ocean. *Proceedings of the National Academy of Sciences*, 116(14):6647–6652, 2019.

[243] Norman R Pace. A molecular view of microbial diversity and the biosphere. *Science*, 276(5313):734–740, 1997.

[244] P Pamilo and M Nei. Relationships between gene trees and species trees. *Molecular biology and evolution*, 5(5):568–583, 1988.

[245] Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290, 2004.

[246] Donovan H Parks, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature biotechnology*, 36(10):996–1004, 2018.

[247] Donovan H Parks, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, 25(7):1043–1055, 2015.

[248] Nicholas D. Pattengale, Krister M. Swenson, and Bernard M E Moret. Uncovering Hidden Phylogenetic Consensus. In *Lecture Notes in Computer Science*, volume 6053 LNBI, pages 128–139. 2010.

[249] Talima Pearson, Heidie M Hornstra, Jason W Sahl, Sarah Schaack, James M Schupp, Stephen M Beckstrom-Sternberg, Matthew W O'Neill, Rachael A Priestley, Mia D Champion, James S Beckstrom-Sternberg, Gilbert J Kersh, James E Samuel, Robert F Massung, and Paul Keim. When Outgroups Fail; Phylogenomics of Rooting the Emerging Pathogen, Coxiella burnetii. *Systematic biology*, 62(5):752–762, jul 2013.

[250] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[251] David Penny. Criteria for optimising phylogenetic trees and the problem of determining the root of a tree. *Journal of molecular evolution*, 8(2):95–116, 1976.

[252] Evgeny Perkovsky and Piotr Wegierek. Aphid–buchnera–ant symbiosis; or why are aphids rare in the tropics and very rare further south? *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 107(2-3):297–310, 2016.

[253] Céline Petitjean, Philippe Deschamps, Purificación López-García, and David Moreira. Rooting the domain archaea by phylogenomic analysis supports the foundation of the new kingdom proteoarchaeota. *Genome biology and evolution*, 7(1):191–204, 2015.

[254] Hervé Philippe, Henner Brinkmann, Richard R. Copley, Leonid L. Moroz, Hiroaki Nakano, Albert J. Poustka, Andreas Wallberg, Kevin J. Peterson, and Maximilian J. Telford. Acoelomorph flatworms are deuterostomes related to Xenoturbella. *Nature*, 470(7333):255–258, 2011.

[255] Hervé Philippe, Henner Brinkmann, Dennis V Lavrov, D Timothy J Littlewood, Michael Manuel, Gert Wörheide, and Denis Baurain. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS biology*, 9(3):e1000602, 3 2011.

[256] Hervé Philippe and Jacqueline Laurent. How good are deep phylogenetic trees? *Current Opinion in Genetics & Development*, 8(6):616–623, 1998.

[257] Hervé Philippe, Damien M. de Vienne, Vincent Ranwez, Béatrice Roure, Denis Baurain, and Frédéric Delsuc. Pitfalls in supermatrix phylogenomics. *European Journal of Taxonomy*, 2017.

[258] Noah J. Planavsky, Dan Asael, Axel Hofmann, Christopher T. Reinhard, Stefan V. Lalonde, Andrew Knudsen, Xiangli Wang, Frantz Ossa Ossa, Ernesto Pecoits, Albertus J. B. Smith, Nicolas J. Beukes, Andrey Bekker, Thomas M. Johnson, Kurt O. Konhauser, Timothy W. Lyons, and Olivier J. Rouxel. Evidence for oxygenic photosynthesis half a billion years before the great oxidation event. *Nature Geoscience*, 7(4):283–286, March 2014.

[259] Rafael I Ponce-Toledo, Philippe Deschamps, Purificación López-García, Yvan Zivanovic, Karim Benzerara, and David Moreira. An early-branching freshwater cyanobacterium at the origin of plastids. *Current Biology*, 27(3):386–391, 2017.

[260] David Posada. Selection of models of dna evolution with jm odel t est. In *Bioinformatics for DNA sequence analysis*, pages 93–112. Springer, 2009.

[261] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3):e9490, 2010.

[262] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. FastTree-2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, 5(3):e9490, mar 2010.

[263] Richard O Prum, Jacob S Berv, Alex Dornburg, Daniel J Field, Jeffrey P Townsend, Emily Moriarty Lemmon, and Alan R Lemmon. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature*, 526(7574):569–573, oct 2015.

[264] Pere Puigbò, Yuri I Wolf, and Eugene V Koonin. Search for a'tree of life'in the thicket of the phylogenetic forest. *Journal of biology*, 8(6):1–17, 2009.

[265] Mário J.F. Pulquério and Richard A. Nichols. Dates from the molecular clock: how wrong can we be? *Trends in Ecology & Evolution*, 22(4):180–184, 2007.

[266] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research*, page gks1219, 2012.

[267] Maryam Rabiee and Siavash Mirarab. Forcing external constraints on tree inference using ASTRAL. *BMC Genomics*, 21(S2):218, 4 2020.

[268] Maryam Rabiee and Siavash Mirarab. INSTRAL: Discordance-Aware Phylogenetic Placement Using Quartet Scores. *Systematic Biology*, 69(2):384–391, 8 2020.

[269] Andrew Rambaut, Alexei J Drummond, Dong Xie, Guy Baele, and Marc A Suchard. Posterior summarization in bayesian phylogenetics using tracer 1.7. *Systematic biology*, 67(5):901–904, 2018.

[270] Andrew Rambaut and Nicholas C. Grass. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3):235–238, 06 1997.

[271] Andrew Rambaut and Edward Holmes. The early molecular epidemiology of the swine-origin a/h1n1 human influenza pandemic. *PLoS currents*, 1:RRN1003–RRN1003, Aug 2009.

[272] Hemalatha Golaconda Ramulu, Mathieu Groussin, Emmanuel Talla, Remi Planel, Vincent Daubin, and Céline Brochier-Armanet. Ribosomal proteins: toward a next generation standard for prokaryotic systematics? *Molecular phylogenetics and evolution*, 75:103–117, 2014.

[273] Bruce Rannala and Ziheng Yang. Inferring Speciation Times under an Episodic Molecular Clock. *Systematic Biology*, 56(3):453–466, 06 2007.

[274] Kasie Raymann, Céline Brochier-Armanet, and Simonetta Gribaldo. The two-domain tree of life is linked to a new root for the Archaea. *Proceedings of the National Academy of Sciences*, 112(21):6670–6675, 2015.

[275] Anna-Louise Reysenbach and Everett Shock. Merging genomes with geochemistry in hydrothermal ecosystems. *Science*, 296(5570):1077–1082, 2002.

[276] Christian Rinke, Patrick Schwientek, Alexander Sczyrba, Natalia N. Ivanova, Iain J. Anderson, Jan-Fang Cheng, Aaron Darling, Stephanie Malfatti, Brandon K. Swan, Esther A. Gies, Jeremy A. Dodsworth, Brian P. Hedlund, George Tsiamis, Stefan M. Sievert, Wen-Tso Liu, Jonathan A. Eisen, Steven J. Hallam, Nikos C. Kyrpides, Ramunas Stepanauskas, Edward M. Rubin, Philip Hugenholtz, and Tanja Woyke. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459):431–437, July 2013.

[277] David F Robinson and Leslie R Foulds. Comparison of phylogenetic trees. *Mathematical biosciences*, 53(1-2):131–147, 1981.

[278] DF Robinson and LR Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2):131–147, 1981.

[279] Sebastien Roch and Sagi Snir. Recovering the Treelike Trend of Evolution Despite Extensive Lateral Genetic Transfer: A Probabilistic Analysis. *Journal of Computational Biology*, 20(2):93–112, 2013.

[280] Andrew J Roger, Sergio A Muñoz-Gómez, and Ryoma Kamikawa. The origin and diversification of mitochondria. *Current Biology*, 27(21):R1177–R1192, 2017.

[281] Jeffrey A. Rosenfeld, Ansel Payne, and Rob DeSalle. Random roots and lineage sorting. *Molecular phylogenetics and evolution*, 64(1):12–20, 2012.

[282] Greg W Rouse, Nerida G Wilson, Jose I Carvajal, and Robert C Vrijenhoek. New deep-sea species of Xenoturbella and the position of Xenacoelomorpha. *Nature*, 530(7588):94–97, feb 2016.

[283] Frank Rutschmann. Molecular dating of phylogenetic trees: A brief review of current methods that estimate divergence times. *Diversity & Distributions*, 12(1):35–48, 1 2006.

[284] A Rzhetsky and M Nei. Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular Biology and Evolution*, 10(5):1073–1095, 09 1993.

[285] Pavel Sagulenko, Vadim Puller, and Richard A Neher. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution*, 4(1), 01 2018. vex042.

[286] N Saitou and M Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 07 1987.

[287] Leonidas Salichos and Antonis Rokas. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, 497(7449):327–31, 2013.

[288] Andreas Sand, Morten Holt, Jens Johansen, Rolf Fagerberg, Gerth Brodal, Christian Pedersen, and Thomas Mailund. Algorithms for Computing the Triplet and Quartet Distances for Binary and General Trees. *Biology*, 2(4):1189–1209, 9 2013.

[289] Andreas Sand, Morten K. Holt, Jens Johansen, Gerth Stølting Brodal, Thomas Mailund, and Christian N. S. Pedersen. tqDist: a library for computing the quartet and triplet distances between binary or general trees. *Bioinformatics*, 30(14):2079–2080, 03 2014.

[290] Michael J. Sanderson. A Nonparametric Approach to Estimating Divergence Times in the Absence of Rate Constancy. *Molecular Biology and Evolution*, 14(12):1218–1231, 12 1997.

[291] Michael J Sanderson. Estimating rate and time in molecular phylogenies: beyond the molecular clock? In *Molecular systematics of plants II*, pages 242–264. Springer, 1998.

[292] Michael J. Sanderson. Estimating Absolute Rates of Molecular Evolution and Divergence Times: A Penalized Likelihood Approach. *Molecular Biology and Evolution*, 19(1):101–109, 1 2002.

[293] Michael J. Sanderson. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, 19(2):301–302, 1 2003.

[294] Michael J. Sanderson and Amy C. Driskell. The challenge of constructing large phylogenetic trees, 2003.

[295] Brice AJ Sarver, Matthew W Pennell, Joseph W Brown, Sara Keeble, Kayla M Hardwick, Jack Sullivan, and Luke J Harmon. The choice of tree prior and molecular clock does not substantially affect phylogenetic inferences of diversification rates. *PeerJ*, 7:e6334, 2019.

[296] Aaron M Satkoski, Nicolas J Beukes, Weiqiang Li, Brian L Beard, and Clark M Johnson. A redox-stratified ocean 3.2 billion years ago. *Earth and Planetary Science Letters*, 430:43–53, 2015.

[297] Erfan Sayyari and Siavash Mirarab. Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. *Molecular biology and evolution*, 33(7):1654–1668, jul 2016.

[298] Erfan Sayyari, James B Whitfield, and Siavash Mirarab. Fragmentary Gene Sequences Negatively Impact Gene Tree and Species Tree Reconstruction. *Molecular Biology and Evolution*, 34(12):3279–3291, 12 2017.

[299] Erfan Sayyari, James B Whitfield, and Siavash Mirarab. Discovista: Interpretable visualizations of gene tree discordance. *Molecular Phylogenetics and Evolution*, 122:110–115, 2018.

[300] Bettina E Schirrmeister, Patricia Sanchez-Baracaldo, and David Wacey. Cyanobacterial evolution during the precambrian. *International Journal of Astrobiology*, 15(3):187–204, 2016.

[301] Jeffrey H. Schwartz and Bruno Maresca. Do molecular clocks run at all? a critique of molecular systematics. *Biological Theory*, 1(4):357–371, Dec 2006.

[302] The scikit-bio development team. scikit-bio: A bioinformatics library for data scientists, students, and developers, 2020.

[303] Nicola Segata, Daniela Börnigen, Xochitl C Morgan, and Curtis Huttenhower. Phylophlan is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature communications*, 4(1):1–11, 2013.

[304] R Shankarappa, J B Margolick, S J Gange, A G Rodrigo, D Upchurch, H Farzadegan, P Gupta, C R Rinaldo, G H Learn, X He, X L Huang, and J I Mullins. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *Journal of virology*, 73(12):10489–502, 12 1999.

[305] Xing-xing Shen, Chris Todd Hittinger, and Antonis Rokas. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution*, 1(5):0126, 4 2017.

[306] Patrick M Shih. Photosynthesis and early earth. *Current Biology*, 25(19):R855–R859, 2015.

[307] Patrick M Shih, James Hemp, Lewis M Ward, Nicholas J Matzke, and Woodward W Fischer. Crown group oxyphotobacteria postdate the rise of oxygen. *Geobiology*, 15(1):19–29, 2017.

[308] Patrick M Shih, Lewis M Ward, and Woodward W Fischer. Evolution of the 3-hydroxypropionate bicycle and recent transfer of anoxygenic photosynthesis into the chloroflexi. *Proceedings of the National Academy of Sciences*, 114(40):10749–10754, 2017.

[309] Le Si Quang, Olivier Gascuel, and Nicolas Lartillot. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*, 24(20):2317–2323, 2008.

[310] BW Silverman. Density estimation for statistics and data analysis. In *Monographs on Statistics and Applied Probability*, number 1951, page 176. Chapman & Hall/CRC, London, 1986.

[311] Mark P Simmons and John Gatesy. Coalescence vs. concatenation: sophisticated analyses vs. first principles applied to rooting the angiosperms. *Molecular phylogenetics and evolution*, 91:98–122, may 2015.

[312] Sagi Snir, Tandy Warnow, and Satish Rao. Short quartet puzzling: A new quartet-based phylogeny reconstruction algorithm. *Journal of Computational Biology*, 15(1):91–103, 2008. PMID: 18199023.

[313] Sagi Snir, Yuri I. Wolf, and Eugene V. Koonin. Universal Pacemaker of Genome Evolution. *PLoS Computational Biology*, 8(11):e1002785, 11 2012.

[314] Sen Song, Liang Liu, Scott V Edwards, and Shaoyuan Wu. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences*, 109(37):14942–7, sep 2012.

[315] Rochelle M Soo, James Hemp, Donovan H Parks, Woodward W Fischer, and Philip Hugenholtz. On the origins of oxygenic photosynthesis and aerobic respiration in cyanobacteria. *Science*, 355(6332):1436–1440, 2017.

[316] Julien Soubrier, Mike Steel, Michael SY Lee, Clio Der Sarkissian, Stéphane Guindon, Simon YW Ho, and Alan Cooper. The influence of rate heterogeneity among sites on the time dependence of molecular rates. *Molecular biology and evolution*, 29(11):3345–3358, 2012.

[317] Anja Spang, Eva F Caceres, and Thijs JG Ettema. Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. *Science*, 357(6351):eaaf3883, 2017.

[318] Mark S. Springer and John Gatesy. The gene tree delusion. *Molecular phylogenetics and evolution*, 94(Part A):1–33, jul 2016.

[319] Mark S. Springer and John Gatesy. On the importance of homology in the age of phylogenomics. *Systematics and Biodiversity*, 16(3):210–228, 4 2018.

[320] Alexandros Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 08 2006.

[321] Alexandros Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

[322] Alexandra Stechmann and Thomas Cavalier-Smith. Rooting the eukaryote tree by using a derived gene fusion. *Science*, 297(5578):89–91, 2002.

[323] Jeffrey W. Streicher, James A. Schulte, and John J Wiens. How Should Genes and Taxa be Sampled for Phylogenomic Analyses with Missing Data? An Empirical Study in Iguanian Lizards. *Systematic Biology*, 65(1):128–145, 1 2016.

[324] C Strömpl, BJ Tindall, H Lünsdorf, TY Wong, ER Moore, and H Hippe. Reclassification of clostridium quercicolum as dendrosporobacter quercicolus gen. nov., comb. nov. *International journal of systematic and evolutionary microbiology*, 50(1):101–106, 2000.

[325] Daniel J. Stynes, George L. Peterson, and Donald H. Rosenthal. Log transformation bias in estimating travel cost models. *Land Economics*, 62(1):94–103, 1986.

[326] Marc A Suchard, Philippe Lemey, Guy Baele, Daniel L Ayres, Alexei J Drummond, and Andrew Rambaut. Bayesian phylogenetic and phylodynamic data integration using beast 1.10. *Virus evolution*, 4(1):vey016, 2018.

[327] Jeet Sukumaran and Mark T Holder. DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, 26(12):1569–1571, 2010.

[328] G J Szöllõsi, E Tannier, Vincent Daubin, and Bastien Boussau. The inference of gene trees with species trees. *Systematic Biology*, 64(1):e42–e62, 7 2014.

[329] Shinichi Takaichi, Takashi Maoka, Kazuto Takasaki, and Satoshi Hanada. Carotenoids of gemmatimonas aurantiaca (gemmatimonadetes): identification of a novel carotenoid, deoxyoscillol 2-rhamnoside, and proposed biosynthetic pathway of oscillol 2, 2‘-dirhamnoside. *Microbiology*, 156(3):757–763, 2010.

[330] Hideyuki Tamaki, Yasuhiro Tanaka, Hiroaki Matsuzawa, Mizuho Muramatsu, Xian-Ying Meng, Satoshi Hanada, Kazuhiro Mori, and Yoichi Kamagata. Armatimonas rosea gen. nov., sp. nov., of a novel bacterial phylum, armatimonadetes phyl. nov., formally called the candidate phylum op10. *International journal of systematic and evolutionary microbiology*, 61(6):1442–1447, 2011.

[331] Koichiro Tamura, Fabia Ursula Battistuzzi, Paul Billing-Ross, Oscar Murillo, Alan Filipski, and Sudhir Kumar. Estimating divergence times in large molecular phylogenies. *Proceedings of the National Academy of Sciences*, 109(47):19333–19338, 11 2012.

[332] Koichiro Tamura, Qiqing Tao, and Sudhir Kumar. Theoretical Foundation of the RelTime Method for Estimating Divergence Times from Variable Evolutionary Rates. *Molecular Biology and Evolution*, 35(7):1770–1782, 03 2018.

[333] Ge Tan, Matthieu Muffato, Christian Ledergerber, Javier Herrero, Nick Goldman, Manuel Gil, and Christophe Dessimoz. Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. *Systematic Biology*, 64(5):778–791, 9 2015.

[334] R. Tarrío, F. Rodríguez-Trelles, and F. J. Ayala. Tree rooting with outgroups when they differ in their nucleotide composition from the ingroup: the Drosophila saltans and willistoni groups, a case study. *Molecular phylogenetics and evolution*, 16(3):344–349, 2000.

[335] James E Tarver, Mario dos Reis, Siavash Mirarab, Raymond J Moran, Sean Parker, Joseph E. O'Reilly, Benjamin L King, Mary J. O'Connell, Robert J Asher, Tandy Warnow, Kevin J Peterson, Philip C.J. Donoghue, and Davide Pisani. The Interrelationships of Placental Mammals and the Limits of Phylogenetic Inference. *Genome Biology and Evolution*, 8(2):330–344, feb 2016.

[336] Simon Tavaré. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86, 1986.

[337] R Development Core Team. R: A Language and Environment for Statistical Computing, 2011.

[338] Jeffrey L. Thorne and Hirohisa Kishino. Divergence time and evolutionary rate estimation with multilocus data. *Systematic Biology*, 51(5):689–702, September 2002.

[339] Jeffrey L. Thorne, Hirohisa Kishino, and Ian S. Painter. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution*, 15(12):1647–1657, 12 1998.

[340] Thu-Hien To, Matthieu Jung, Samantha Lycett, and Olivier Gascuel. Fast Dating Using Least-Squares Criteria and Algorithms. *Systematic Biology*, 65(1):82–97, 1 2016.

[341] Akiko Tomitani, Andrew H Knoll, Colleen M Cavanaugh, and Terufumi Ohno. The evolutionary diversification of cyanobacteria: molecular–phylogenetic and paleontological perspectives. *Proceedings of the National Academy of Sciences*, 103(14):5442–5447, 2006.

[342] Fernando Domingues Kümmel Tria, Giddy Landan, and Tal Dagan. Phylogenetic rooting using minimal ancestor deviation. *Nature Ecology & Evolution*, 1(1):0193, 2017.

[343] Pranjal Vachaspati and Tandy Warnow. ASTRID: Accurate Species TRees from Internode Distances. *BMC Genomics*, 16(Suppl 10):S3, 2015.

[344] William SJ Valdar. Scoring residue conservation. *Proteins: structure, function, and bioinformatics*, 48(2):227–241, 2002.

[345] Yoshiki Vazquez-Baeza, Meg Pirrung, Antonio Gonzalez, and Rob Knight. EMPeror: a tool for visualizing high-throughput microbial community data. *Gigascience*, 2(1):2047–217X, 2013.

[346] EM Volz and SDW Frost. Scalable relaxed clock phylogenetic dating. *Virus evolution*, 3(2), 2017.

[347] Erik M. Volz, Katia Koelle, and Trevor Bedford. Viral Phylodynamics. *PLoS Computational Biology*, 9(3), 2013.

[348] Huai Chun Wang, Bui Quang Minh, Edward Susko, and Andrew J. Roger. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Systematic Biology*, 67(2):216–235, 2018.

[349] Tandy Warnow, Bernard M. E. Moret, and Katherine St. John. Absolute convergence: True trees from short sequences. In *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '01, page 186–195, USA, 2001. Society for Industrial and Applied Mathematics.

[350] Joel O. Wertheim, Mathieu Fourment, and Sergei L. Kosakovsky Pond. Inconsistencies in Estimating the Age of HIV-1 Subtypes Due to Heterotachy. *Molecular Biology and Evolution*, 29(2):451–456, 2 2012.

[351] Joel O. Wertheim, Michael J. Sanderson, Michael Worobey, and Adam Bjork. Relaxed Molecular Clocks, the Bias–Variance Trade-off, and the Quality of Phylogenetic Inference. *Systematic Biology*, 59(1):1–8, 1 2010.

[352] Kristi M. Westover, Joseph P. Rusinko, Jon Hoin, and Matthew Neal. Rogue taxa phenomenon: A biological companion to simulation analysis. *Molecular Phylogenetics and Evolution*, 69(1):1–3, 2013.

[353] Travis J Wheeler and John D Kececioglu. Multiple alignment by aligning alignments. *Bioinformatics*, 23(13):i559–i568, 2007.

[354] William B Whitman. The modern concept of the procaryote. *Journal of Bacteriology*, 191(7):2000–2005, 2009.

[355] N. J. Wickett, Siavash Mirarab, Nam Nguyen, Tandy Warnow, Eric J Carpenter, Naim Matasci, Saravanaraj Ayyampalayam, M. S. Barker, J Gordon Burleigh, Matthew A. Gitzendanner, Brad R Ruhfel, E. Wafula, J. P. Der, S. W. Graham, S. Mathews, Michael Melkonian, Douglas E Soltis, Pamela S Soltis, N. W. Miles, C. J. Rothfels, L. Pokorny,

A. J. Shaw, L. DeGironimo, Dennis W Stevenson, B. Surek, J. C. Villarreal, B. Roure, Hervé Philippe, Claude W DePamphilis, T. Chen, Michael K Deyholos, R. S. Baucom, Toni M Kutchan, M. M. Augustin, Jian Wang, Y. Zhang, Z. Tian, Z. Yan, X. Wu, X. Sun, G. K.-S. Wong, and Jim Leebens-Mack. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences*, 111(45):E4859–4868, oct 2014.

[356] Tom A Williams, Peter G Foster, Cymon J Cox, and T Martin Embley. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*, 504(7479):231–236, 2013.

[357] Tom A Williams, Sarah E Heaps, Svetlana Cherlin, Tom M W Nye, Richard J Boys, and T Martin Embley. New substitution models for rooting phylogenetic trees. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678):20140336, sep 2015.

[358] Tom A Williams, Gergely J Szöllősi, Anja Spang, Peter G Foster, Sarah E Heaps, Bastien Boussau, Thijs J G Ettema, and T Martin Embley. Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proceedings of the National Academy of Sciences*, 114(23):E4602–E4611, 2017.

[359] Carl R Woese and George E Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090, 1977.

[360] Carl R Woese, Otto Kandler, and Mark L Wheelis. Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579, 1990.

[361] Matthias Wolf, Tobias Müller, Thomas Dandekar, and J Dennis Pollack. Phylogeny of firmicutes with special reference to mycoplasma (mollicutes) as inferred from phosphoglycerate kinase amino acid sequence data. *International journal of systematic and evolutionary microbiology*, 54(3):871–875, 2004.

[362] Kelly C. Wrighton, Brian C. Thomas, Itai Sharon, Christopher S. Miller, Cindy J. Castelle, Nathan C. VerBerkmoes, Michael J. Wilkins, Robert L. Hettich, Mary S. Lipton, Kenneth H. Williams, Philip E. Long, and Jillian F. Banfield. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science*, 337(6102):1661–1665, September 2012.

[363] Yufeng Wu. A practical method for exact computation of subtree prune and regraft distance. *Bioinformatics*, 25(2):190, 2009.

[364] Zhenxiang Xi, Liang Liu, and Charles C. Davis. Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. *Molecular phylogenetics and evolution*, 92(3):63–71, November 2015.

[365] Zhenxiang Xi, Liang Liu, Joshua S Rest, and Charles C Davis. Coalescent versus concatenation methods and the placement of Amborella as sister to water lilies. *Systematic biology*, 63(6):919–932, 11 2014.

[366] Xuhua Xia. DAMBE7: New and Improved Tools for Data Analysis in Molecular Biology and Evolution. *Molecular Biology and Evolution*, 35(6):1550–1552, 6 2018.

[367] Xuhua Xia and Qun Yang. A distance-based least-square method for dating speciation events. *Molecular Phylogenetics and Evolution*, 59(2):342–353, 2011.

[368] Xiao Xiao, Ethan P. White, Mevin B. Hooten, and Susan L. Durham. On the use of log-transformation vs. nonlinear regression for analyzing biological power laws. *Ecology*, 92(10):1887–1894, 2011.

[369] Ying Xu and Nicolas Glansdorff. Was our ancestor a hyperthermophilic procaryote? *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, 133(3):677–688, 2002.

[370] Ziheng Yang. A heuristic rate smoothing procedure for maximum likelihood estimation of species divergence times. *Dong wu xue bao.[Acta Zoologica Sinica]*, 50(4), 2004.

[371] Pablo Yarza, Pelin Yilmaz, Elmar Pruesse, Frank Oliver Glöckner, Wolfgang Ludwig, Karl-Heinz Schleifer, William B Whitman, Jean Euzéby, Rudolf Amann, and Ramon Rosselló-Móra. Uniting the classification of cultured and uncultured bacteria and archaea using 16s rrna gene sequences. *Nature Reviews Microbiology*, 12(9):635–645, 2014.

[372] John Yin, Chao Zhang, and Siavash Mirarab. ASTRAL-MP: Scaling ASTRAL to very large datasets using randomization and parallelization. *Bioinformatics*, 35(20):3961–3969, 2019.

[373] Anne D Yoder and Ziheng Yang. Estimation of primate speciation dates using local molecular clocks. *Molecular Biology and Evolution*, 17(7):1081–1090, 2000.

[374] Yun Yu, Tandy Warnow, and Luay Nakhleh. Algorithms for MDC-based multi-locus phylogeny inference: beyond rooted binary gene trees on single alleles. *Journal of Computational Biology*, 18(11):1543–1559, nov 2011.

[375] Natalya Yutin and Michael Y. Galperin. A genomic update on clostridial phylogeny: Gram-negative spore formers and other misplaced clostridia. *Environmental Microbiology*, pages n/a–n/a, July 2013.

[376] Natalya Yutin, Pere Puigbò, Eugene V Koonin, and Yuri I Wolf. Phylogenomics of prokaryotic ribosomal proteins. *PloS one*, 7(5):e36972, 2012.

[377] Katarzyna Zaremba-Niedzwiedzka, Eva F. Caceres, Jimmy H. Saw, Disa Bäckström, Lina Juzokaite, Emmelien Vancaester, Kiley W. Seitz, Karthik Anantharaman, Piotr Starnawski, Kasper U. Kjeldsen, Matthew B. Stott, Takuro Nunoura, Jillian F. Banfield, Andreas

Schramm, Brett J. Baker, Anja Spang, and Thijs J. G. Ettema. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, 541(7637):353–358, January 2017.

[378] Chao Zhang, Maryam Rabiee, Erfan Sayyari, and Siavash Mirarab. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19(S6):153, 5 2018.

[379] Chao Zhang, Yiming Zhao, Edward Louis Braun, and Siavash Mirarab. TAPER: Pinpointing errors in multiple sequence alignments despite varying rates of evolution. *bioRxiv*, page 2020.11.30.405589, 2020.

[380] Xue Zhang, Bo Tu, Li-rong Dai, Paul A Lawson, Zhen-zhen Zheng, Lai-Yan Liu, Yu Deng, Hui Zhang, and Lei Cheng. Petroclostridium xylanilyticum gen. nov., sp. nov., a xylan-degrading bacterium isolated from an oilfield, and reclassification of clostridial cluster iii members into four novel genera in a new hungateiclostridiaceae fam. nov. *International journal of systematic and evolutionary microbiology*, 68(10):3197–3211, 2018.

[381] Xiaofan Zhou, Xing-Xing Shen, Chris Todd Hittinger, and Antonis Rokas. Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. *Molecular biology and evolution*, 35(2):486–503, 2018.

[382] Verena Zimorski, Chuan Ku, William F Martin, and Sven B Gould. Endosymbiotic theory for organelle origins. *Current opinion in microbiology*, 22:38–48, 2014.

[383] Yuanqiang Zou, Wenbin Xue, Guangwen Luo, Ziqing Deng, Panpan Qin, Ruijin Guo, Haipeng Sun, Yan Xia, Suisha Liang, Ying Dai, Daiwei Wan, Rongrong Jiang, Lili Su, Qiang Feng, Zhuye Jie, Tongkun Guo, Zhongkui Xia, Chuan Liu, Jinghong Yu, Yuxiang Lin, Shanmei Tang, Guicheng Huo, Xun Xu, Yong Hou, Xin Liu, Jian Wang, Huanming Yang, Karsten Kristiansen, Junhua Li, Huijue Jia, and Liang Xiao. 1, 520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nature Biotechnology*, 37(2):179–185, February 2019.

[384] Emile Zuckerkandl. Molecular disease, evolution, and genetic heterogeneity. *Horizons in biochemistry*, pages 189–225, 1962.

[385] Derrick J. Zwickl, Joshua C. Stein, Rod A. Wing, Doreen Ware, and Michael J. Sanderson. Disentangling methodological and biological sources of gene tree discordance on Oryza (Poaceae) chromosome 3. *Systematic Biology*, 63(5):645–659, 2014.