ASSESSMENT OF TAXON SAMPLING ON PHYLOGENETIC RECONSTRUCTIONS AND TIMETREES: A NEW METHODOLOGY AND APPLICATION

by

CHRISTOPHER LOWELL EDWARD POWELL

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY IN BIOLOGICAL SCIENCE

2023

Oakland University

Rochester, Michigan 48309

Doctoral Advisory Committee:

Fabia Ursula Battistuzzi, Ph.D. Chair

Taras K. Oleksyk, Ph.D.

Sara E. Blumer-Schuette, Ph.D.

To my wife, Samantha, my mother Suzanne, and my father Gary

ACKNOWLEDGEMENTS

I am deeply grateful to Dr. Fabia Battistuzzi for taking me into her lab knowing that I traversed an ecological path and had minimal knowledge on early earth prokaryotes and genomics/proteomics. Her mentorship has sharpened my interest and passion for a multitude of fields I knew little about. I appreciate her guidance and consistent support throughout this whole process. I couldn't have asked for a better mentor on my journey. I would like to express my gratitude to my committee members, Dr. Taras Oleksyk and Dr. Sara Blumer-Schuette for their time and helpful insights in the writing of my dissertation, proposal, and dissertation defense. Honestly, thank you so much. You were an amazing committee.

Additionally, I would like to thank my loving family and wife who have been the support structure and motivation that got me through. Thank you to my mother Suzanne, my father Gary, my wife Samantha, my grandmother Elaine, my uncle Mark, and my aunt Fran. Thank you all for putting up with my perpetual schooling and craziness throughout the years.

Lastly, I would like to thank Oakland University Biology department and everyone who has helped me throughout the many years. Thank you specifically to Dr. Thomas Raffel, Dr. Keith Berven, and Gary Miller for everything they did for me throughout the years. Thank you to Jan Bills, Kathy Lesich, and Cathy Starnes for all of the work they do to keep the department rolling forward!

ABSTRACT

ASSESSMENT OF TAXON SAMPLING ON PHYLOGENETIC RECONSTRUCTIONS AND TIMETREES: A NEW METHODOLOGY AND APPLICATION

by

Christopher Lowell Edward Powell

Over the past three decades, computational capabilities have grown at such a rapid rate that they have given rise to many computationally heavy science fields such as phylogenomics. As increasingly more genomes are sequenced in the three domains of life, larger and more species-complete phylogenetic tree reconstructions are leading to a better understanding of the Tree of Life and the evolutionary histories in deep times. However, these large datasets pose unique challenges from a modeling and computational perspective: accurately describing the evolutionary process of thousands of species is still beyond the capability of current evolutionary models while the computational burden limits our ability to exhaustively explore and test multiple hypotheses. These limitations become even more problematic when attempting to estimate the absolute times within these phylogenetic reconstructions (timetrees). These time estimations are not only constrained computationally by run times and resource requirements but also bound by the availability of fossil data to estimate divergence times for the evolution of species

(primary calibrations). All of these issues are particularly severe in prokaryotes, because of the high number of species available in databases, their large evolutionary variability, and the few primary calibrations available. Yet, they represent two out of the three domains of life and are therefore key to reconstructing the Tree of Life. This combination of computational and data constraints is forcing researchers to make choices on the datasets being analyzed without a clear understanding of the consequences of these choices on the accuracy of the results obtained. This work presents an in-depth analysis of the effects of dataset choices on the reconstruction of phylogenetic histories using a newly developed tool (Phylogenetic Assessment of Taxon Sampling) that will enable fast, simple, and reproducible testing of taxon sampling. The PATS pipeline is available on GitHub: https://github.com/BlabOaklandU/PATS

TABLE OF CONTENTS

ACKNOWLEDGMENTS iv

ABSTRACT v

LIST OF TABLES x

LIST OF FIGURES xi

LIST OF ABBREVIATIONS xiv

CHAPTER ONE

BACKGROUND AND SIGNIFICANCE 1

1.1: Introduction to Phylogenetics 1

CHAPTER TWO

TESTING FOR PHYLOGENETIC

INSTABILITY VIA TAXON SAMPLING 5

2.1: Introduction 5

2.2: Topology instability: common causes and solutions 6

2.2.1: Choice of phylogenetic method 6

2.2.2: Modeling the evolutionary process 9

2.3: The case of taxon sampling 10

2.4: Methodology 12

2.4.1: Reconstruction of species phylogenies 13

2.4.2: Phylogenetic assessment of taxon sampling 15

2.5: The PATS Pipeline Deconstructed 18

TABLE OF CONTENTS - CONTINUED

2.5.1: Fasta Standardizer 19

2.5.2: Orthology Determination 20

2.5.2.1: BLAST 20

2.5.2.2: Protein Ortho 22

2.5.3: Selection of orthologs 23

2.5.4: Alignment, gap filtering, and concatenation 24

2.5.5: Tree reconstruction 24

2.6: Taxon sampling scenarios 25

2.6.1: Keep One 25

2.6.2: Remove One 26

2.6.3: Remove Group 26

2.7: Phylogenetic comparisons 27

2.8: PATS optional features 27

2.8.1: Archiving Functionality 27

2.8.2: Multi-node BLASTing 28

2.9: Results 28

2.10: Conclusions 36

CHAPTER THREE

TESTING PRIMARY VS SECONDARY CALIBRATIONS

IN A SIMULATED ENVIRONMENT FOR

MOLECULAR CLOCK APPLICATIONS 40

3.1 Introduction 40

3.2: Methodology 44

3.2.1: Dataset 44

3.2.2: Time estimation 46

3.2.3: Measure of accuracy 47

TABLE OF CONTENTS – CONTINUED

3.3 Results 48

3.3.1: Assessing accuracy of primary calibrations 49

3.3.2: Assessing accuracy of secondary calibrations 52

3.4: Conclusions 57

CHAPTER FOUR

TESTING THE EFFECT OF SPECIES CHOICE

ON TIME TREE RECONSTRUCTIONS AND

ITS IMPLEMENTATION INTO THE PATS PIPELINE 61

4.1: Introduction 61

4.2.1: Implementation of RelTime into PATS 64

4.2: Results 65

4.2.1: Branch lengths 65

4.2.2: Divergence times 66

CHAPTER FIVE

CONCLUSIONS 73

REFERENCES 75

APPENDIX 96

LIST OF TABLES

Table 1	Shows the bootstrap values for all of the KeepOne scenarios with respect to the standard (full) tree of 252 species.	34
Table 2	Scenarios of varying uncertainty around the true time (TT) for primary calibration boundaries.	49
Table 3	Average slope of estimated times *vs.* simulated true times (ET accuracy) for ten 30K concatenations.	51
Table 4	Confidence intervals (CIs) accuracy (proportion of CIs that include the simulated true time) for each of the seven scenarios.	53
Table 5	Confidence interval (CI) precision relative to the simulated true time.	56
Table 1A	Actinobacteria species that were a part of the 103 dataset. *N/A depicts entry that is no longer available in NCBI database.*	96
Table 2A	Actinobacteria species that were a part of the 252 Dataset	99

LIST OF FIGURES

Figure 1	Growth of a phylogenetic approach from its initial proposition to its widespread application in many fields.	1
Figure 2	Growth trends of published studies in the phylogenetics field.	3
Figure 3	Maximum Likelihood landscape plot	9
Figure 4	The total number of fully sequenced genomes in multiple Bacteria groups from NCBI as of 9 March 2023.	11
Figure 5	Major steps in a phylogenetic reconstruction analysis	14
Figure 6	Relation of different phylogenetic tree reconstruction Approaches.	15
Figure 7	Schematic representation of the three types of analyses that can be completed with the PATS pipeline.	18
Figure 8	Topologies and bootstrap supports of major groups in a phylogeny of 104 Actinobacteria.	31
Figure 9	Topologies comparison between Keep One and full tree of the 252 Actinobacteria.	31

LIST OF FIGURES – CONTINUED

Figure 10	Topologies used in simulated analyses for Tree A (A) and Tree B (B).	44
Figure 11	Empirically derived parameters used to simulate alignments.	46
Figure 12	Comparison of average slopes for each of the seven scenarios in the four-calibration setting.	50
Figure 13	Simulated true times *vs.* relative times from RelTime before calibrations are applied.	57
Figure 14	Topologies and branch length supports of major groups in a phylogeny of 103 Actinobacteria.	66
Figure 15	Tree topologies for each of the full tree, the Nocardia KeepOne tree, and the Rhodococcus KeepOne tree.	67
Figure 16	Divergence times plot estimated via RelTime for each Keepone scenario and full tree analysis.	69
Figure 17	Divergence time plot estimated via RelTime for each calibrated node.	70
Figure 18	Divergence time plot estimated via RelTime for each permutation scenario versus the full trees.	72

LIST OF FIGURES – CONTINUED

Figure A1	ET accuracy (estimated time *vs.* simulated true time) for tree A with three primary calibrations.	106
Figure A2	ET accuracy (estimated time *vs.* simulated true time) for tree B with one secondary calibrations.	107
Figure A3	ET accuracy (estimated time *vs.* simulated true time) for tree AB with three distant primary calibrations.	108
Figure A4	ET accuracy (estimated time *vs.* simulated true time) for tree AB with three distant primary calibrations were only data points for Tree B are plotted.	109
Figure A5	ET accuracy (estimated time *vs.* simulated true time) for Tree B with one primary calibration.	110
Figure A6	Plotted confidence interval ranges for each node of the KeepOne permutation against the full tree that contains the same calibrations.	111

LIST OF ABBREVIATIONS

PATS: Phylogenetic assessment of taxon sampling

ToL: Tree of Life

Evo: Evolutionary

Biochem: Biochemistry

Mol: Molecular

Bio: Biology

Multidis: Multidisciplinary

Sci: Science

Biotech: Biotechnology

Micro: Microbiology

Fw: Freshwater

Con: Conservation

Env: Environmental

Res: Research

Math: Mathematical

Comp: Computational

Pharmaco: Pharmacology

Pharm: Pharmacy

App: Applied

Chem: Chemistry

Med: Medicine

Morph: Morphology

DNA: Deoxyribonucleic acid

RNA: Ribonucleic acid

GTR: Generalized time-reversable

ML: Maximum likelihood

UPGMA: Unweight pair group method

MCMC: Markov Chain Monte Carlo

RAxML: Randomized Axelerated Maximum Likelihood

RBS: Rapid bootstrapping

LIST OF ABBREVIATIONS – CONTINUED

JC: Jukes-Cantor

LG: Le-Gascuel

LG + G: Le-Gascuel + gamma rate

LG4X: Le-Gascuel four-matrix model + gamma rate heterogeneity

LG4M: Le-Gascuel four-matrix model + gamma rate heterogeneity

JTT: Jones-Taylor-Thornton

WAG: Whelan and Goldman

AIC: Akaike information criterion

CAT: Categorized model

BLAST: Basic local alignment search tool

MSP: Maximal segment per measure

MUSCLE: Multiple sequence comparison by log-expectation

AU: Approximately unbiased

SH: Shimodaira-Hasegawa

NCBI: National center for biotechnology and information

HKY: Hasegawa-Kishno-Yano

MYA: Million years ago

HPCC: High performance computational cluster

iCer: Institute for cyber-enabled research

ET: Estimated time

TT: True time

CI: Confidence interval

M-N/R: Mycobacterium – Nocardia/Rhodococcus

S-T: Segniliparus – Tsukamurella

Vs: vers

CHAPTER ONE

BACKGROUND AND SIGNIFICANCE

1.1: Introduction to Phylogenetics

Ever since the first phylogenetic tree was drawn by Charles Darwin in his famous book 'On the Origin of Species' (Fig. 1A), the visual representation of species-to-species connections (branches) started to become common throughout biology and, eventually, blossomed into the fields of phylogenetics and phylogenomics [1, 2]. The idea of connecting species based upon shared characteristics provides a powerful comparative framework to uncover the evolutionary strategies of organisms driven by genetic drift and natural selection. These strategies are now applied across many different fields of biology as a single unifying principle to interpret biological evolution (Fig. 1B).

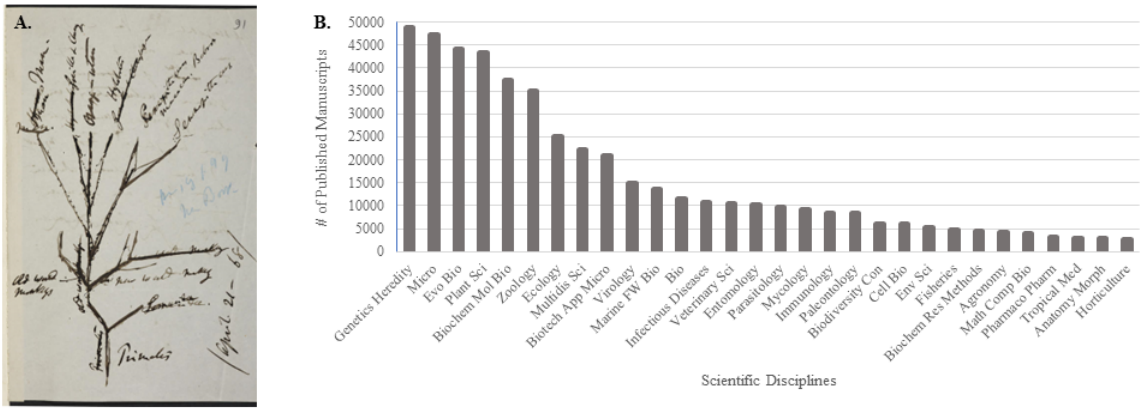


Figure 1: Growth of a phylogenetic approach from its initial proposition to its widespread application in many fields. **(A.)** Charles Darwin's species-tree sketch that is thought to be one of the first phylogenetic trees created **(B.)** The top thirty science fields that have publications that contain the word 'phylogen' (from Web of Science). Evo:evolutionary; biochem:biochemistry; mol:molecular; bio:biology; multidis:multidisciplinary; sci:science; biotech:biotechnology; micro:microbiology; fw:freshwater; con:conservation; env:environmental; res:research; math:mathematical; comp:computational; pharmaco:pharmacology; pharm:pharmacy; app:applied; chem:chemistry; med:medicine; morph:morphology

This comparative framework has also revolutionized the methodologies used in biology, especially since its application to genetic and genomic data. Instead of measuring relatedness via few and taxonomically constrained morphological characteristics, it is now more common to evaluate relatedness at the genetic level (DNA, proteins), effectively connecting all life forms. Moreover, this molecular phylogenetic approach expands the size of datasets by orders of magnitude, thus minimizing the number of false relations that parallel [3] and convergent evolution [4] can spuriously create when only few morphological characteristics are considered [5–8]. On the one hand, from a morphological standpoint, it might seem that if two species share one (or even a few) highly similar characteristics they might also share a recent common ancestor [5, 9]. However, using many genes can, and often does, reveal that the observed similarity is driven by functional adaptation rather than recent genetic inheritance, thus revealing a falsely identified recent common ancestor [10]. On the other hand, unicellular prokaryotes cannot be efficiently compared to other prokaryotes and multicellular eukaryotes using morphology alone due to the lack of extensive morphological features [11]. Thus, owing to its shared nature, genetic data has allowed the generation of large phylogenetic (Fig. 2A) and phylogenomic (Fig. 2B) trees, including Tree of Life (ToL) reconstructions [12–14], that are mapping the evolutionary histories of all known species [15–18]. In the past 15 years, technological advances have favored the use of phylogenomic ("big data") approaches to reconstruct the history of life allowing detailed representations of all domains, especially the microbial ones (in particular Bacteria).

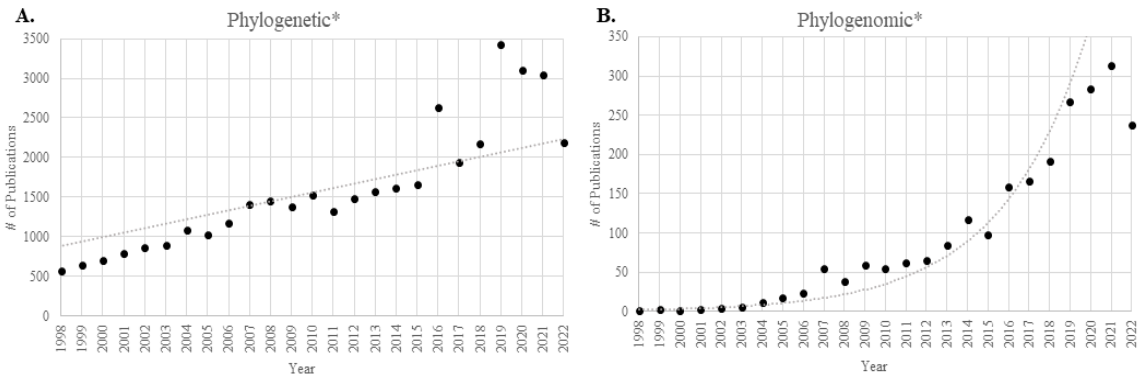


Figure 2: Growth trends of published studies in the phylogenetics field. Studies that contain the word phylogenetic **(A.)** show a linear trend but phylogenomic studies **(B.)** are exponentially growing [from the Web of Science].

A phylogenetic tree can include three main components: 1) topology, 2) branch lengths, and 3) time (relative/absolute). There are many ways to attempt to optimize these three main components, but the consensus has been that the use of more data produces more accurate results [15, 16, 18–20]. However, this assumption is problematic for two reasons: *first*, the presence of systematic biases can negatively affect the accuracy of the results obtained especially in the presence of large datasets [21]; *second*, computational constraints often prevent the use of "all data" and lead to sub-setting of information. Thus, even in the "genomics age", a deeper understanding of how dataset choices affect phylogenetic reconstructions is fundamental. This exploration should happen at multiple levels: the starting point is seeing the effect of species choice on the topology of the generated tree, which allows to test the idea of 'representative species', the selected few that are to represent an entire clade. After the topology has been optimized, next would be the comparison of branch lengths that represent the relative proportion of genetic change (i.e., longer branch lengths equate to larger amounts of genetic change). Branch lengths are used to generate relative times and can, therefore, provide information on

evolutionary rates among lineages. Lastly, the calculation of time scaled to absolute values using calibration points (times of known species origin) should also be evaluated in light of dataset choices made.

The rationale behind species choice testing is that reconstructions of phylogenetic and phylogenomic trees are dependent on an accurate modeling of complex evolutionary processes that can be affected by many species-specific characteristics. These are often sources of "phylogenetic noise" which can mask or alter the "true" phylogenetic signal [21–23]. These noise sources can be classified under two broad categories: methodology and data-driven. The methodology-driven sources include the type of dataset (e.g., nucleotide *vs.* amino acids), the tree reconstruction framework (e.g., maximum likelihood, Bayesian), and the complexity of the evolutionary model used (e.g., GTR) [24]. The data-driven sources, instead, include dataset size, (e.g., number of genes or species), compositional biases (the preferential use of some nucleotides/amino acids over others), and evolutionary rates [24–26]. The difficulty of accounting for these issues is that, in most cases, they are intertwined with effects that are caused by more than one source at a time. An example of this synergy is taxon sampling, generally defined as the strategy used to select a set of taxa to reconstruct an evolutionary history of a larger group. The total number of taxa used affects the overall dataset size and the choice of taxa can have repercussions on compositional biases, evolutionary rates, and the evolutionary models chosen.

CHAPTER TWO

TESTING FOR PHYLOGENETIC INSTABILITY VIA TAXON SAMPLING

2.1: Introduction

The rise in availability of phylogenomic-level data initially suggested that more species and more sites would lead to more accurate and stable phylogenetic tree reconstructions because of the larger amount of information available to model the evolutionary processes [25, 27]. While this expectation was shown to be valid for some datasets, for others systematic biases reinforced by large datasets became a prominent issue with many phylogenetic relationships still being uncertain even when utilizing all available data [e.g.,15, 17, 18, 26–28]. At the same time, these large-scale analyses also expose data and computational limitations that negatively affect our ability to use the full complement of species available. For example, to reduce missing data it is often necessary to exclude species with a highly reduced genome and smaller datasets are required for specific analyses (e.g., Bayesian reconstructions) because of the large computational burden [31–33]. Thus, in many cases, some level of taxon sampling is still required and an evaluation of the consequences of these choices on the stability of phylogenies would be advisable. However, this evaluation is rarely performed because of the time and computational resources it requires. Indeed, comparisons of published reconstructed Trees of Life from different studies show that they produce topologically different trees and that these differences are often attributed to changes in dataset size and methodology but without robust evidence supporting any one of the topologies over the

others [12–14, 34]. Thus, researchers are often left wondering which topology is more accurate, if any. This question can be addressed within a simulated framework to test the relative effects of taxon sampling and other sources of bias on the phylogenetic reconstruction process. A recent study took this approach and showed that discrepancies in the topology of deep nodes of *Terrabacteria* are caused by taxon sampling even when other parameters are held constant [21].

2.2: Topology instability: common causes and solutions

2.2.1: Choice of phylogenetic method There are two main approaches that are used to estimate phylogenetic trees: character-based and distance-based. The character-based approaches utilize either Bayesian or frequentist inference, which includes maximum likelihood (ML) and maximum parsimony (MP). Bayesian inference utilizes Bayes' theorem, which relies on the conditional probability of an event [35]. Based on this, a posterior probability of a hypothesis (phylogeny) being true is calculated using (i) the likelihood of an event occurring given the data provided and (ii) a prior of the probability of the hypothesis being true without considering the observed data. This is different from the frequentist (ML) inference that is based upon the idea that an event has an innate probability (theoretical probability) of occurring and that replication of the event will expose this probability [36, 37]. Unlike frequentist inference, Bayesian relies on a prior which can be interpreted as how probable the hypothesis is before accounting for the observed data. This prior is one of the main defining differences between frequentist and Bayesian inference and often a source of debates as the information used to determine the prior can be subjective. Distance approaches are very different from the

character-based ones as they utilize pairwise comparisons of sequences to calculate distance matrices between each pair based on the ratio of sites that differ from each other relative to the total length of the alignment (basic uncorrected distance) [11, 38]. As an example, two distance-based methodologies include neighbor joining [39] and the Unweighted Pair Group Method using arithmetic Averages (UPGMA) [11, 40]. Both the character- and distance-based approaches have strengths and weaknesses and are chosen based on the dataset being explored considering multiple factors such as the expected amount of divergence, the size of the dataset, and the expected computational time. Within the character-based approaches, Bayesian is the slowest due to its utilization of Metropolis-Hastings Markov Chain Monte Carlo (MCMC) inference or a variation of it [41–44]. Maximum likelihood (ML) uses varying maximum likelihood estimators for phylogenetic tree estimation and is significantly faster than Bayesian MCMC inference, especially when using methods specifically designed for large datasets [45–49]. Maximum parsimony is faster than maximum likelihood because it utilizes the parsimony score, the smallest number of substitutions needed for each site summed, which is quicker as it does not depend on a stochastic optimization process [50]. However, Maximum Parsimony is also less accurate, especially for datasets with large evolutionary distances. The distance-based approaches, are also fast and often are utilized in tandem with another approach such as ML or Bayesian as they quickly produce initial tree estimates and can be used to test parameters for models of evolution [11].

Beyond the nature of the algorithm, a multitude of factors play into the run-time of a phylogenetic analysis. One of these major factors is whether the phylogenetic tree algorithm uses a heuristic (best guess) or an exhaustive (exact) approach. An exhaustive

search will evaluate the complete parameter space (trees) whereas a heuristic search will focus on increasingly higher probability paths ignoring those that fall below the highest score (e.g., ML score) at each step. While exhaustive searches are the only ones guaranteed to produce the best scoring phylogeny given the data, they are too time consuming for large datasets, leading most to use the heuristic approach. Due to the nature of a heuristic search that auto-optimizes via a stochastic process, it is possible that it can get stuck in a local maximum which would not represent the best tree topology [51, 52] (Fig. 3). One way to limit this issue is via the process of bootstrapping [53]. Bootstrapping allows for retesting of branch locations by resampling the original dataset and reconstructing a phylogeny for every pseudo replicate, which is very time consuming. However, a rapid bootstrapping (RBS) method has been implemented into RAxML (Randomized Axelerated Maximum Likelihood) to minimize bootstrapping times for tree reconstructions [47, 54]. The RBS method can be applied to the starting tree on the original alignment and facilitates the optimization method before the maximum likelihood is even calculated [47, 54]. The consensus in phylogenetics is for at least 100 bootstrap replicates whenever possible [47]. Another strategy to identify phylogenies that may not be optimal is to run multiple identical analyses and compare the results. If the phylogenies differ, it is likely that the heuristic method got stuck on a local maximum in either one, many, or all of the identical runs.

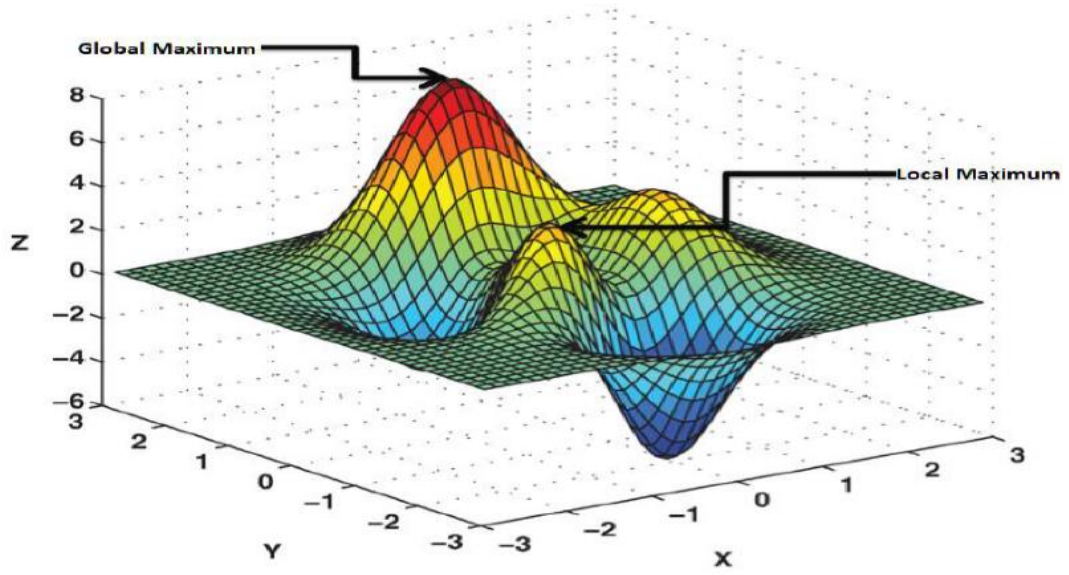


Figure 3: Maximum Likelihood landscape plot. 'Blind' random walks around the X-Y plane attempt to locate the highest peak in the z-plane (global maximum = best tree topology). However, there is a chance of false-positive results if the search algorithm gets stuck on a smaller peak (local maximum).

2.2.2: Modeling the evolutionary process Complex evolutionary processes are modeled considering multiple aspects, primarily substitution probabilities and rate variation among lineages, within lineages, and among sites [55, 56]. Substitution models assign frequencies (weights) to site changes based on their expected probability of occurrence. These models are widely different in complexity starting from two of the simplest ones, Jukes-Cantor (JC) [57] and Dayhoff model [58], for nucleotides and amino acids respectively, to the more complex General Time Reversible (GTR) [59] and Le-Gascuel (LG) models [60]. The JC model is implemented under the assumptions of equal nucleotide frequencies and equal substitution rates whereas the GTR model is used when varying nucleotide frequencies and multiple substitution rates are expected [61]. The LG4X/LG4M, a variation of the LG model, uses different substitution models for the

different evolutionary rate categories via multiple-matrices (single-level mixture models) and can generate higher AIC (Akaike information criterion) values than single-matrix model such as LG, JTT, and WAG [62]. Each of these models can also be associated with a distribution to model non-uniform evolutionary rates at different positions of an alignment (e.g., gamma distribution) [55, 56] and also models to account for changes over time of these parameter (heterotachy) [63]. For these cases, models such as CAT (Bayesian approach) or protein mixture (Maximum Likelihood approach) are available [60, 64].

Finally, while choice of substitution models and phylogenetic methods are the most common decisions that have to be made to perform a phylogenetic analysis, many other aspects of a dataset can affect its accurate evolutionary reconstruction, such as alignment quality, compositional biases, and determination of orthology, among many others. For a review of these methods, see [29, 56].

2.3: The case of taxon sampling

A controversial issue in topology instability is taxon sampling [65–69]. Taxon sampling is the process of selecting species to represent taxonomic groups in a phylogenetic analysis. It can be comprehensive, when every species with available data for a given group is used, or selective, when only representative species are chosen. By definition, taxon sampling is dependent upon the availability of species to choose from, which, in turn, depends on how intensively a taxonomic group is being sequenced (either complete genome sequencing or selected genes). There is a large disparity in sequencing efforts among groups which leads to large datasets being unbalanced in their taxonomic

representation, even when a comprehensive taxon sampling approach is used [70, 71](Fig. 4).

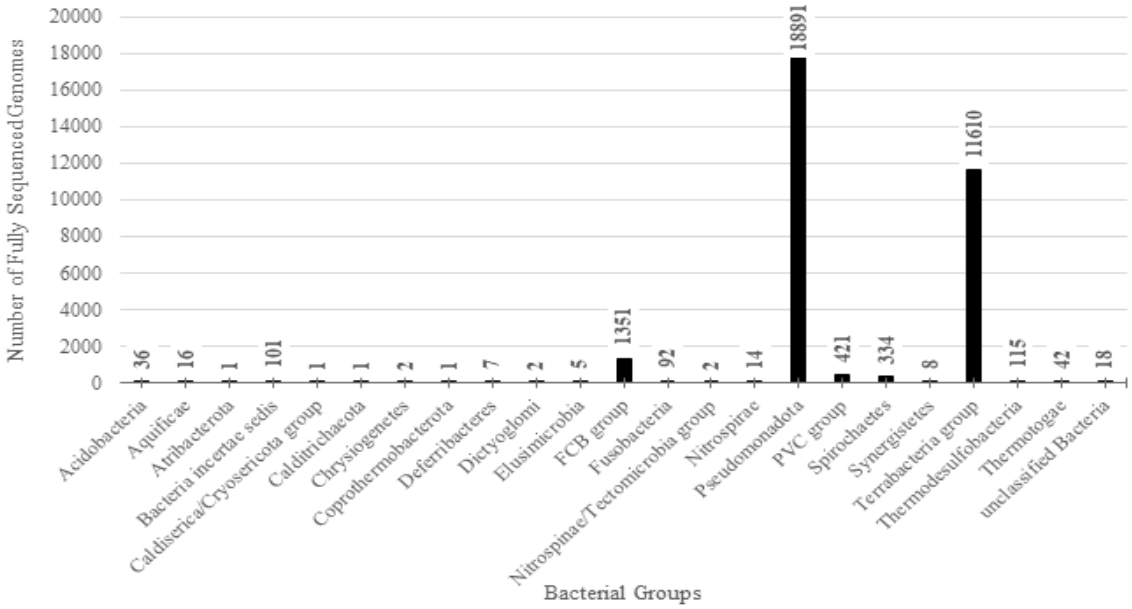


Figure 4: The total number of fully sequenced genomes in multiple Bacteria groups from NCBI as of 9 March 2023.

Thus, under the category “taxon sampling” there are multiple co-occurring issues that can affect the accuracy of a reconstructed phylogeny: (i) the number of species selected; (ii) the identity of the species selected; (iii) and the unbalance among groups that these choices (or data availability) can lead to. Any observed effect is, in reality, not a direct consequence of the process of taxon sampling but an indirect effect of the fact that each chosen (or available) species has unique genome properties and evolutionary processes that can differentially impact the success of the evolutionary modelling approaches we use. As large datasets become increasingly common in phylogenetics, the task of taxon sampling becomes even more difficult and cumbersome. The best way of avoiding taxon sampling would seem to be to include all species known [72–74].

However, the feasibility of this approach is often not achievable or desirable because it causes other limitations in dataset size (the issue of missing data mentioned above) and because run-times of many analyses would become unrealistically long (e.g., Bayesian tree reconstructions). Additionally, even this approach would not solve the issue of an unbalanced tree caused by uneven sequencing. Thus, taxon sampling is a reality of any phylogenetic analysis that needs to be considered and carefully evaluated.

Unfortunately, while testing the effect of taxon sampling is important to determine the accuracy of a phylogeny, the extensive computational time and resources required to do so means, that in most cases, it is not performed. A strategy to mitigate these issues is to use automated computational pipelines that allow fast and easy testing with minimal manual input from a researcher. This is the approach we take in this aim.

2.4: Methodology

The reconstruction of phylogenetic trees is a multi-step process that relies on choices among many approaches and methods. While there is a consensus on the overall sequence of steps that leads to a tree, unique strategies are often applied to specific datasets. The development of multiple software has made the need for studies that benchmark and compare the performance of methods to each other fundamental [32, 33, 75–78]. However, often there is no single method that is overall more accurate or efficient compared to others and choices can be subjective. This subjectivity makes it even more important to assess phylogenetic stability in light of the choices made, since data and methods can affect the final phylogenetic outcome.

2.4.1: Reconstruction of species phylogenies……..Phylogenetic tree reconstructions start with a list of species of interest (here we focus on species trees, but similar protocols are followed for gene trees). From this list, the steps involved in the reconstruction are fairly standard although there are many tools available to perform each step (Fig. 5A). The choice of tools is often subjective but aspects such as computational time and resources required by each software usually play an important role in the decision, especially when large datasets are being analyzed. The first step is to identify shared orthologs among species using tools like OrthoMCL [79], OrthoMCL-DB [80], Ensembl 2018 [81], ProteinOrtho [82], InParanoid 8 [83], eggNOG 5.0 [84], and OrthoFinder [85]. All these software are based on the principle of sequence similarity but they have unique strengths and weaknesses. For example, ProteinOrtho is known to be fast and flexible compared to other tools that are too slow to work with large data sets (OrthoMCL) and/or rely on prebuilt databases based on specific species and groups (Ensembl, OrthoMCL-DB, InParanoid, and eggNOG). Once orthologs are identified, the second step is to align them to identify sites that share an evolutionary history. This can be done with a multiple sequence alignment software such as T-Coffee [86], MUSCLE [87], ClustalOmega [88], or MAFFT [89]. All these methods are widely used although T-Coffee has been shown to be slower and less accurate than MUSCLE and ClustalOmega [86–88] and thus it may be less applicable to large datasets. After multiple alignments are created, the **t**hird step is to choose between a supermatrix [90] or a supertree [91] reconstruction strategy. The supermatrix strategy is most commonly used because it increases the amount of information available to the phylogenetic method. The supermatrix consists of the concatenation of each aligned gene into a long, single

sequence. Finally, this supermatrix becomes the input for a phylogenetic tree reconstruction software such as FastTree [45, 46], RAxML [47], PAML (BASEML / CODEML) [92], and IQTREE [48, 49], which all rely on maximum likelihood estimation. For Bayesian tree building software some of the most popular ones include BEAST [42, 44], BUCKy [93], MrBayes [43], PAML (MCMCTREE) [92], BATWING [94], and PATIS [95]. For maximum parsimony tree building software there are MPBoot [50], TNT [96], MEGA [97–99], and PAUP [100]. As for the distance-based approach, MEGA [97–99], QuickTree [101] and IQPNNI [102] can all utilize neighbor-joining method. MEGA can also implement the UPGMA method (and also Maximum Likelihood approaches) (Fig. 6). Software differ in terms of functionality, availability of built-in options and applicability to large datasets and, in many cases, the choice of one or the other is rather subjective. Recently, IQTree has become increasingly more common because of its flexibility in number of options for tree reconstruction and testing [48, 49].

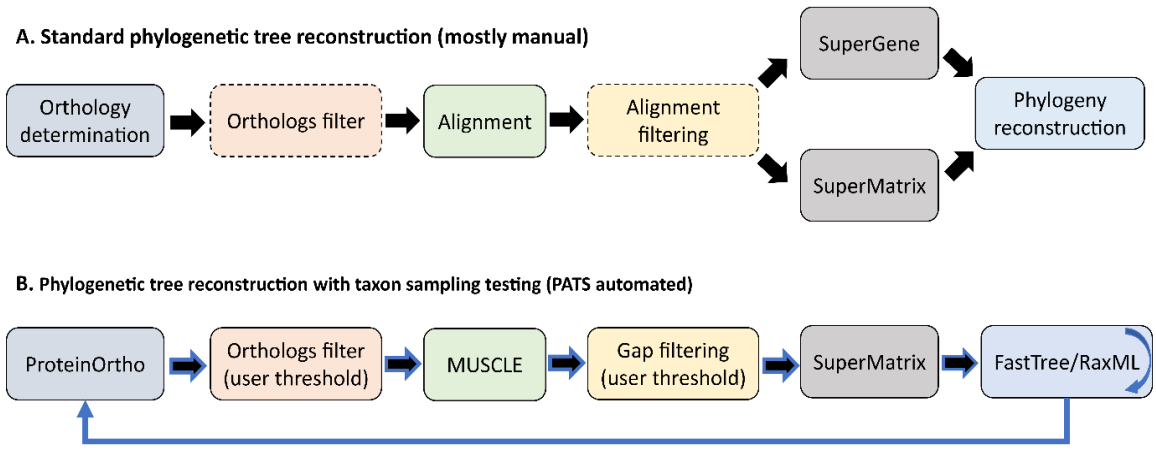


Figure 5: Major steps in a phylogenetic reconstruction analysis. **(A.)** A step-by-step flow chart of the major steps in a phylogenetic tree reconstruction analysis. Each block represents a unique step, and each black arrow represents a manual step for the user. **(B.)** Details of the phylogenetic process implemented in PATS. All these steps are automated and controlled by a single configuration file. The blue arrows represent the implementation of the taxon sampling feature of the PATS pipeline. Taxon sampling can be done within the tree creation step '*poPipe_rerunTreesOnly*' or restarting from the orthology determination step '*poPipe_rerunFullAnalysis*'. Dotted borders: facultative steps.

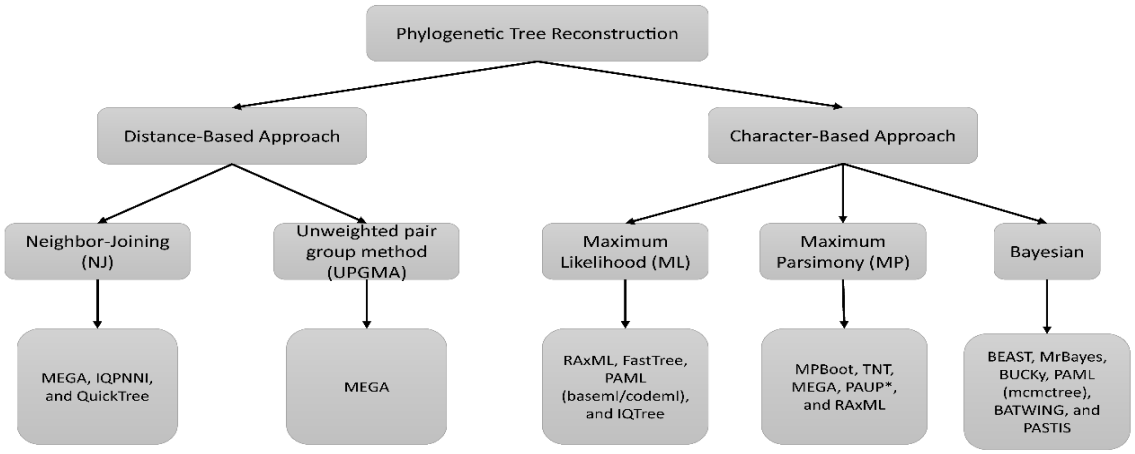


Figure 6: Relation of different phylogenetic tree reconstruction approaches. Commonly used software for each of the approaches are shown. *(Modified from Sleator 2011).*

For many phylogenetic studies, the reconstruction of the tree is the final step in the process followed by evolutionary, ecological, or molecular interpretations of the topology and branch lengths obtained. However, as mentioned above, often the reconstructed trees are influenced by the choices made throughout the process and, therefore, before interpretations are made, the phylogeny should be tested for its stability in light of parameter permutations. In other cases, phylogenetic tree reconstruction is followed by timetree estimations, which also relies heavily on the accuracy of the topology obtained. In these cases, very little is known of the effect of parameter permutations on the estimation of relative and absolute times, especially in cases when the topology does not change.

2.4.2: Phylogenetic assessment of taxon sampling We focus on the effect of taxon sampling on phylogenetic stability for two reasons: *first*, it encompasses many other methodologies and data-driven biases and, *second*, because it is an issue rarely addressed in phylogenetic analyses. Within microbial phylogenetics, taxon sampling can

be especially challenging because of the large numbers of species sequenced, at least in some taxonomic groups, and because of the large variation within genera. For example, within the relatively small *Deinococcus-Thermus* phylum, the GC content of scaffold and fully sequenced genomes varies from ~45% to >70% suggesting that compositional biases driven by the choice of species could play an important role in the accuracy of the phylogenetic placement of this group. Here, we discuss the development and use of a computational pipeline, Phylogenetic Assessment of Taxon Sampling (PATS), that allows users to explore the effects of taxon sampling and species choice on phylogenetic tree reconstructions in a fast and efficient way. This pipeline is unique in its functionality and it is designed to work with large datasets of protein or nucleotide data, although, generally, protein data is the optimal type of data for ToL and deep phylogenies. PATS is therefore particularly useful for the study of deep divergences in the history of life. With a series of permutation scenarios, PATS integrates the exploration of number of species and choice of species: users can provide a list of species and choose to iteratively remove a single species from it (*RemoveOne*), remove all species except one (*KeepOne*), or remove groups of species (*RemoveGroup*). These permutations allow for an automated way to test multiple hypothesis-driven scenarios in which different sets of species are used to estimate a phylogenetic history. PATS is a native Linux pipeline that combines the consensus steps of a typical phylogenetic analysis with additional functionalities (Fig. 5B). It is user-friendly and has an archiving feature to allow its application to large datasets. The PATS pipeline implements the fastest software for phylogenetic reconstruction (from orthology determination to tree building) in a single package that gives users full control to design the type of taxon sampling test to perform (Fig. 7). This

includes a single hypothesis test (standard run, Fig. 7A), a multiple hypothesis test where the number of orthologous genes remains the same, but species change (standard run + species permutation with tree reconstruction, Fig. 7B), and a multiple hypothesis testing where orthologous genes are redetermined for each tree (species permutations with orthology redetermination, Fig. 7C). PATS is composed of 21 bash scripts and 18 Perl scripts that work in chain to drive the pipeline from start to finish. Of the steps depicted in (Fig. 7), step 2, step 4, step 5, and step 7 were specifically developed for this pipeline. These are the orthology Filter script that takes in a user set threshold and filters the output from ProteinOrtho, the gap filter that removes a threshold of gaps from the alignment file, the concatenation filter that generates the supermatrix for the tree generation software, and lastly the permutations that allow for KeepOne / RemoveOne / Remove Group. The permutation step is the core of PATS as it allows for two unique ways of taxon sampling, within tree permutation and orthology redetermination. A few of the other scrips are cleanup (FileCleanUp.pl, Archive.pl, and ArchiveCleanUp.pl) and file checking scrips (fastaStandardize.sh, fastaStandardizer.pl, and fastaStandardizerCheck.pl). Many of the remaining scripts are specific to the type of taxon sampling settings the user chooses before the pipeline runs. The PATS pipeline is available on our GitHub: https://github.com/BlabOaklandU/PATS. Below we discuss the theory and application of PATS.

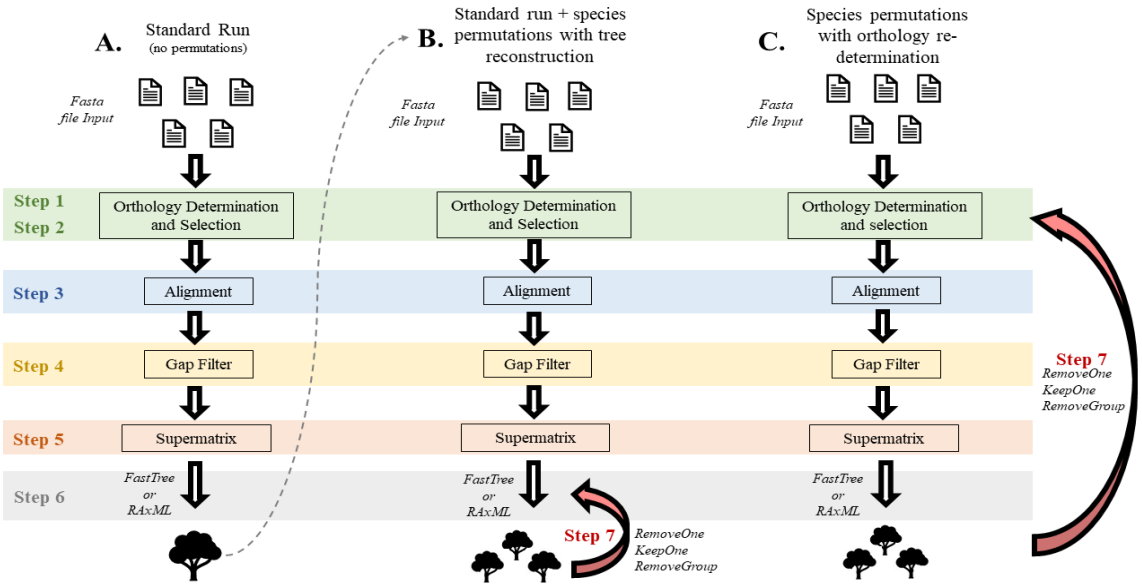


Figure 7: Schematic representation of the three types of analyses that can be completed with the PATS pipeline. **(A)** A standard run that allows for the reconstruction of a single phylogenetic tree based on the species provided; **(B)** A standard run with species permutations at the tree reconstruction stage. This produces multiple phylogenetic trees. Before this pathway can be used, the standard run (pathway **(A)**) needs to be completed; **(C)** A standard run with the implementation of permutations at the orthology determination step. This pathway will generate multiple tree outputs (based on the permutation selected) and can be run as a standalone analysis, without **(A)**.

2.5: The PATS Pipeline Deconstructed

Phylogenetic tree reconstructions have become very popular as computational capabilities have allowed the analysis of increasingly larger datasets. Indeed, reconstructing the Tree of Life has become one of the most sought-after puzzles in the world of phylogenetics and many studies have produced ToL using hundreds of genes shared by known and environmental (unclassified) lineages [12–14, 103–105].

However, because of variable gene loss, one of the first choices to be made in ToL (and other phylogenetic) reconstructions is whether or not to increase the number of orthologous genes by excluding some species (taxon sampling). For this reason, ToL reconstructions vary significantly in number of species used, number of orthologous

genes used, and total number of sites. Moreover, even when an identical dataset is used, ToL reconstructions still suffer from an unbalanced representation of taxonomic groups because of unequal sequencing efforts in different lineages. For example, the phylum *Deinococcus-Thermus* (*Deinococcota*), which includes many extremophilic bacteria, has a little over 80 species fully sequenced whereas *Proteobacteria* (*Pseudomonadota*), which includes some important human pathogens, has a close to 19,000. Within PATS, choices that determine number of species, number of orthologs, and permutation scenarios for taxon sampling are controlled by a configuration file (poPipeSettings.sh) that the user will edit before the start of an analysis.

2.5.1: Fasta Standardizer……..Before the first major step of the PATS pipeline is called, the fasta standardizer runs. This reads through all of the fasta files in the main fasta directory, checking for any data anomalies and errors. It will check for proper file type (.fasta) as well as the type of data in each fasta file, either genomic (nucleotide) or proteomic (amino acid). If any discrepancy is found, the pipeline will error and quit, preventing a significant amount of wasted time. Next, the fasta standardizer will perform some text formatting to remove discrepancies in file naming or other content. For example, it will replace special symbols (- * ‘ ( ) [ ] { }) and spaces with an underscore ‘_’ in the fasta file name. To visualize one of the outputs of ProteinOrtho, the fasta standardizer also modifies each descriptive line inside each file to ensure they are of the same length (by adding “X” symbols). This does not affect the analysis but simplifies visualization in later steps. Backups of the original fasta files (unedited) and standardized fasta files (edited) are generated. If running multiple iterations of the pipeline, the fasta standardizer will check for accession length across all files and if they are equal, it will

bypass the data editing portion of the fasta standardizer script. Lastly, multiple log files are generated for the users. One useful generated log (*dataTypeLog.txt*) contains a breakdown of the data type within each fasta file as well as a detailed amino acid or nucleotide count for the file as a whole. A few of the other logs are outputted for user debugging purposes.

2.5.2: Orthology Determination The first major step in the PATS pipeline is orthology determination. Orthologous genes arise via a speciation event and can be traced back to a recent common ancestor based on genetic and, often, functional similarity [82]. Orthologs are used because they represent the history of species, contrary to paralogs that represent the history of individual genes within species through gene duplications and losses. Although for simplicity we focus on species trees, the procedures, and requirements, to test phylogenetic stability are the same in gene trees. This step is completed using either the default parameters for the Basic Local Alignment Search Tool (BLAST+) or DIAMOND [106, 107] for either proteins or translated DNA and setting the connectivity (PoPipe_proteinOrthoConn) for ProteinOrtho within the configuration file. For utilization of BLAST+ within PATS, blastn, blastp, blastx, and autoblast are all available settings.

2.5.2.1: BLAST To accurately and quickly estimate orthologs, BLAST is paired with ProteinOrtho. BLAST can be used to quickly match a query sequence to one contained within a sequence database [108]. BLAST was introduced in 1990 and, to the current day, is still being updated and maintained due to popularity and quick outputs [108–110]. BLAST's main step, called maximal segment pair measure (MSP), is used to

obtain all combinations of pairwise comparisons between sequences. The MSP is defined as the highest scoring pair of identical length segments chosen from two sequences [109, 110]. MSP implements the use of substitution matrices (PAM [58], BLOSUM [111]) to compare each residue, or monomer, of a sequence. These matrices are generated by the number of unique characters of interest in the string, i.e., for amino acids it is a 20 unique character alphabet. The numbers of unique characters define the dimension of the matrix, 20 x 20 for amino acids, with each row and column containing a unique character. These matrices are then 'scored' based upon a match (e.g., +5) or a mismatch (e.g., -4) for each site and then summed across all sites [109]. The algorithm then will generate all combinations of scores and find the global optimum for the two sequences being compared. The highest score sequence can be of any length but must be a contiguous stretch of residues [109, 110]. All combinations of species are bi-directionally 'blasted', per sequence, to produce a list of high score sequence matches. BLAST calculates an 'expected value' (e-value) that represents the number of chance alignments from a database. Smaller e-values are assigned to matches with higher similarity (and, thus, lower probability of being a chance match). The computational time of this heuristic evaluation is proportional to the product of the lengths of the two sequences [109, 110].

BLAST+ is the current up-to-date version of the base BLAST that has grown with the demands of large data and ease of usability [108]. BLAST+ adds the ability of the original BLAST heuristic to work on long query sequences by breaking them into smaller segments for processing [108]. BLAST+ can be locally installed on a computer and is part of the PATS pipeline package. For protein sequence query data there is an alternative to BLAST, called DIAMOND, that can be used to generate the pairwise outputs.

DIAMOND is extremely quick at amino acid pairwise comparisons via a double indexing protocol that makes it significantly faster than BLASTP and is implemented as an optional alignment choice [106, 107].

2.5.2.2: Protein Ortho ProteinOrtho uses an extended version of reciprocal alignment heuristic to scan through the blasted sequences for orthologs [82]. Before ProteinOrtho can search for orthologs, BLAST (or DIAMOND) must propagate a full collection of pairwise comparisons as well as the similarity statistics (e-values, p-values, bit scores, and raw scores). The output of this similarity analysis is ranked based upon these statistics and ProteinOrtho views these alignments and combines the high-ranking ones to determine orthologous groups [82]. Multiple algorithms, such as the MCL-algorithm and MultiParanoid algorithm, which finds in-paralogs, are used to determine Clusters of Orthologous Groups (COGs) based on the BLAST/DIAMOND analysis output [82]. First, ProteinOrtho filters the BLAST/DIAMOND results via two criteria: (1) a minimal level of sequence identity must be present in the alignment, and (2) a minimum fraction of the query protein must be covered by the alignment [82]. Then, ProteinOrtho normalizes the scores from the BLAST output so that e-values from different pairwise comparisons can be directly compared. It does this by assigning higher similarities for related proteomes as well as allowing lower similarity for larger evolutionary distances [82]. Lastly, a sparse graph is generated for clustering purposes where each link (edge) represents a connection based on pairs of similar proteins in different species and each vertex represents the proteins in the input [82]. The sensitivity of ProteinOrtho is controlled by the connectivity parameter (PoPipe_proteinOrthoConn) which defaults to 0.1 [82]. Increasing the value of this parameter will decrease the number of orthologs

detected because of removal of co-orthologs. To get the most comprehensive set of COGs (which will likely include co-orthologs), it is best to set the connectivity to 0 and filter the obtained orthologs afterwards (see next step). Currently, two versions of ProteinOrtho are available within the PATS pipeline (v6.0.14 and v6.1.7). Version 6.0.14 is considered a stable version and is the one available for use on Galaxy [112] while version 6.1.7 is the most up-to-date version that is actively being worked on.

2.5.3: Selection of orthologs Once orthologs are identified, a user-defined parameter will determine how much missing data is allowed in the dataset. Missing data is defined as an ortholog present in some species but not others. This can be accomplished with OrthoSpeciesFilter, a script written specifically for PATS to screen the output of ProteinOrtho and remove clusters of orthologs that have fewer species than the threshold set by the user. This threshold is specified before the analysis starts as part of the configuration file (poPipeSettings.sh) with the parameter PoPipe_speciesFilterCutoff. As an example, if a user sets the threshold at 90, all ortholog clusters that have 89% or less of the species represented will be removed from future steps. Additionally, the orthologs filtered by OrthoSpeciesFilter have to satisfy another requirement: being present in single copy. This is important for species trees because paralogs originate via duplication events and, thus, can affect the tree reconstruction process by creating spurious clusters. Within OrthoSpeciesFilter, paralogs are identified in clusters with a number of species lower than the user-defined threshold but with more genes than the number of species. OrthoSpeciesFilter separately identifies co-orthologs as present in clusters that satisfy the user species threshold but that also contain more genes

than number of species. Also, co-orthologs are eliminated from further analysis. A log of all orthologs, paralogs, and co-orthologs is outputted for the user.

2.5.4: Alignment, gap filtering, and concatenation MUSCLE (Multiple Sequence Comparison by Log-Expectation) is the software implemented in PATS to create multiple sequence alignments [87]. MUSCLE will take each cluster from OrthoSpeciesFilter and generate an alignment with default parameters [87]. Once created, the alignment is filtered via the GapFilter option (PoPipe_gapFilterCutoff) in the configuration file. The GapFilter will screen and remove sites that show gaps over a user-defined threshold. For example, if the user sets the GapFilter threshold to 50, any site that contains ≥ 50% gaps will be removed from every alignment. The GapFilter defaults to 25. The same threshold is applied to all alignments. Finally, after each alignment is screened for gaps, the ConcatenationFilter will concatenate all of the alignments into a supermatrix. This step is automatically implemented in PATS and requires no user-defined parameter. The gene tree approach is not implemented in PATS.

2.5.5: Tree reconstruction PATS currently implements FastTree or RAxML to reconstruct phylogenetic trees. FastTree is the preferred method because it is faster than RAxML with similar accuracies [45, 46, 77]. FastTree is currently set to run the LG + G Model [60] for amino acids, which is likely to be appropriate for datasets with deep evolutionary histories. However, FastTree can also be set to utilize either the WAG or GTR model with or without gamma. The other tree generation option is RAxML, which does take longer as it implements a more extensive maximum likelihood approach. Currently, RAxML will automatically detect the type of data, either protein or nucleotide,

and set protein to ‘PROTGAMMAAUTO’ which will automatically find the best amino acid model to use [47] or default to a GTR with gamma when nucleotides are detected. For nucleotide analyses, variants of the GTR model can be specifically chosen (GTRCAT, GTRCATI, ASC_GTRCAT, GTRGAMMA, ASC_GTRGAMMA or GTRGAMMAI). Tree creation will automatically start for each iteration after the alignment creation and filtering steps are completed.

2.6: Taxon sampling scenarios

Taxon resampling is the defining feature of the PATS pipeline (Fig. 7). It allows a user to automatically run multiple permutation scenarios with a set of species of interest. One of the main features allows testing the direct effects on phylogenetic tree topology when species are removed and orthologs are recalculated or by only removing species without ortholog recalculations. Thus, it is also possible to test for increased alignment size as removing species is likely to lead to higher number of orthologs and, therefore, longer supermatrices. The type of permutations allowed in PATS are *Keep One*, *Remove One*, and *Remove Group,* which are specified by the parameters ‘PoPipe_keepOne’, ‘PoPipe_removeOne’, and ‘poPipe_removeGroup*’* in the configuration file. List of species to permutate are specified in the files ‘keepOne’, ‘removeOne’, and ‘removeGroup’ within the ‘permutations’ directory.

2.6.1: Keep One *Keep One* is a permutation that will take a user-defined list of species and iteratively ‘keep one’ of them in the analysis and remove the others. The number of species on the list will determine the number of iterative runs the pipeline will complete before it finishes. As an example, if the list contains 10 unique species, the

analysis will have a total of 10 iterative outputs where each iteration contains one unique species on that list. This type of analysis allows the user to simultaneously reduce the number of species and test the effect of species choice. The "standard" tree with all species is useful to compare the results of each iteration to discover the effects of each species relative to the others but also of using fewer numbers of species. In an ideal case scenario where number and choice of species do not affect the topology, all trees should be concordant.

2.6.2: Remove One The opposite of *Keep One* is the *Remove One* option, which allows a user to keep all species from a user-defined list except for one that is iteratively removed. By removing a single species, this option is unlikely to significantly change the size of the dataset, but it is useful to investigate the effect of species that are suspected to exert an unduly influence on the topology. For example, species with a strong compositional bias could cluster together because of similarity based on this bias rather than evolutionary history. In this case, removing these species one by one could allow the user to identify those that are causing the observed cluster.

2.6.3: Remove Group This scenario is intermediate between the two previous ones and allows to test the effect of subgroups within a clade. In this option, the user will input the list of species to be removed and, therefore, the pipeline will only run once. A likely application of this scenario is to investigate the effect of a genus or higher taxonomic level on a large phylogeny.

2.7: Phylogenetic comparisons:

The outcome of the selected permutation scenario is a set of phylogenies that can be compared to determine the effect of the hypotheses tested. The procedure implemented in Superson *et al.* [21] suggests comparing the phylogeny obtained with the full dataset (the "standard" run) to those obtained with a subset of it. There are multiple aspects of a phylogeny that can be compared such as topology and branch lengths. To compare topologies, a common measure is the Robinson-Fould that scores pairs of topologies based on the number of shared nodes [113]. This measure can be paired to an analysis of branch lengths to determine if, for example, nodes that are not shared are connected to small branch lengths. This scenario would be expected as it shows that there is little information available for the tree reconstruction method to determine the branching pattern. Alternatively, the Shimodaira-Hasegawa (SH) or the Approximately Unbiased (AU) test can be used to obtain a statistical evaluation of how trees in a set are different from each other [114, 115]. However, these tests require the topologies to share the same set of species, thus some manipulation of the trees obtained with PATS is required to compare only the backbone (deep) nodes.

2.8: PATS optional features

2.8.1: Archiving Functionality One of the larger concerns when testing hypotheses in large-scale phylogenetic analyses is the amount of space (hard disk) that is required to store all of the analysis outputs. PATS includes an auto-archiving function that will create a compressed zip (7zip) of the individual PATS run and delete all of the

files generated during that analysis. This allows for storage optimization and to maximize taxon sampling testing in a single run.

2.8.2: Multi-node BLASTing A time-consuming aspects to phylogenetic tree reconstructions is the orthology determination step (BLAST+ or Diamond). If PATS is running on a high-performance computational cluster (HPCC) the user can set the orthology determination step of ProteinOrtho to be 'distributed' to minimize run times. Currently, PATS allows for this orthology determination step to be pushed to either 2, 5, 10 or 20 nodes. Each node will run a proportion of the orthology determination step, speeding it up substantially.

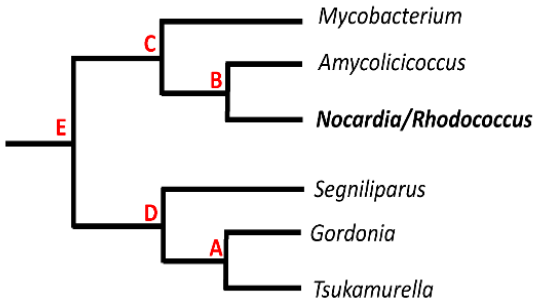
2.9: Results

We utilized a Keep One scenario to test the effect of taxon sampling on two similar subsets of Terrabacteria, more specifically Actinobacteria, that were obtained from NCBI at two different points in times, one in 2015 and one in 2022. The first subset was selected from a previous topology [21] that used only fully sequenced genomes. This subset included 103 Actinobacteria (Table A1). The second subset was downloaded in 2022 and it includes 252 Actinobacteria (Table A2), both with fully sequenced genomes. These two Actinobacteria datasets are a good representation of how quickly genomic/proteomic data has grown in the last seven years. Within the 103 Actinobacteria subset, we identified four *Nocardia* species (*Nocardia nova, N. cyriacigeorgica, N. farcinica, and N. brasiliensis*) and five Rhodococcus species (*Rhodococcus pyridinivorans, R. sp. B7740, R. erythropolis, R. opacus, R. jostii*) as candidate species for this analysis as their placement, and that of closely related clades, in recent, large-

scale phylogenetic tree reconstructions has been unstable [116, 117]. Within the 252 Actinobacteria subset, we identified fifteen Nocardia species (*Nocardia asteroids, N. brasisliensis, N. cyriacigeorgica, N. farcinica, N. gipuzkoensis, N. hauxiensis, N. iowensis, N. mangyaensis, N. otitidiscaviarum_GCF_007362295, N. otitidiscaviarum_GCF_014217765, N. seriolae_GCF_002356035, N. seriolae_GCF_018362975, N. terpenica, N. wallacei, and N. yamanashiensis*), eleven Rhodococcus species (*Rhodococcus aetherivorans, R. coprophilus, R. erythropolis_GCF_000696675, R. erythropolis_GCF_015099595, R. fascians, R. globerulus, R. hoagie, R. koreensis, R. opacus, R. pyridinivorans_GCF_000511305,* and *R. pyridinivorans_GCF_016804345),* and one *Skermania* species (*Skermania piniformis).* First, we obtained a phylogeny with all 103 species (Full tree, Fig. 8A) and a phylogeny with all 252 species (Full tree, Fig. 9A). Then, for the 103 Actinobacteria subset, each Keep One scenario was run from the ortholog determination step to the tree creation step keeping one of the species listed above in the analysis at a time. We set the SpeciesFilter cutoff at 60%, which keeps all orthologous groups that contain 60 – 100% of the species present in the analysis. The chosen threshold was required to increase the number of orthologs while including as many species as possible. One reason for why this cutoff was so low is due to some extremely small proteomes of some of the species in the analyses. With a smaller proteome, it is less likely to match during orthology determination due to the minimal amount of comparable data. However, even with this lenient cutoff, four of our candidate species showed a history of paralogy and were removed from further analysis: *N. farcinica, R. erythropolis, R. jostii, and R.* sp. *B7740*. In addition, we did not analyze results that included *R. pyridinivorans* because it

consistently clustered with a Corynebacterium species in the permutated and full trees. This is an unexpected placement that was not seen in previous studies and suggests the presence of a hidden paralogy or other biases, also supported by its very long branch length [116]. For the 252 Actinobacteria subset, each Keep One scenario was run only from the tree creation step, skipping the ortholog determination step because the SpeciesFilter cutoff was set at 100% which only allows orthologs that are contained by all species present. This threshold was chosen after initially testing at 90% which determined 637 orthologs. We then tested at 99% which determined 329 ortholog, eventually leading us to test 100%. The Keep One scenarios either kept in all of the *Rhodococcus* species and iterated through the *Nocardia* + *Skermania* species (via KeepOne) or kept in all Nocardia + Skermania species and iterated though the *Rhodococcus* species (via KeepOne). No species were removed from the analysis as no odd topologies were found, however, a single Corynebacterium (*Corynebacterium cyclohexanicum*) clustered outside of the rest of the Corynebacterium clade with an outgroup species, *Nukamurella*. This anomaly seems to be linked to a possible misclassification [118] as branch lengths, bootstrap values, and topological location remain relatively constant across all permutations. Paralogy is possible but unlikely due to the number of identified orthologs (142).

A.



B.

KeepOne Species	A	B	C	D	E
Nocardia brasiliensis	91	96	*NA*	37	36
Nocardia cyriacigeorgic	90	92	*NA*	40	50
Nocardia nova	89	99	*NA*	51	51
Rhodococcus opacus	94	89	36	*NA*	33
Full Tree Bootstraps	**86**	**81**	**46**	**27**	**36**

Figure 8*:* Topologies and bootstrap supports of major groups in a phylogeny of 103 Actinobacteria. **(A)** Topology of the genera *Mycobacterium*/*Segniliparus*/*Gordonia*/*Tsukamurella*/*Nocardia*/*Rhodococcus* from the full tree before the KeepOne permutations were applied. Each letter (A – E) represents a node that relates to **(B)** to show the bootstrap values and how the nodes change for each KeepOne scenario. In **(A)** the bolded genera are those used during the permutations.

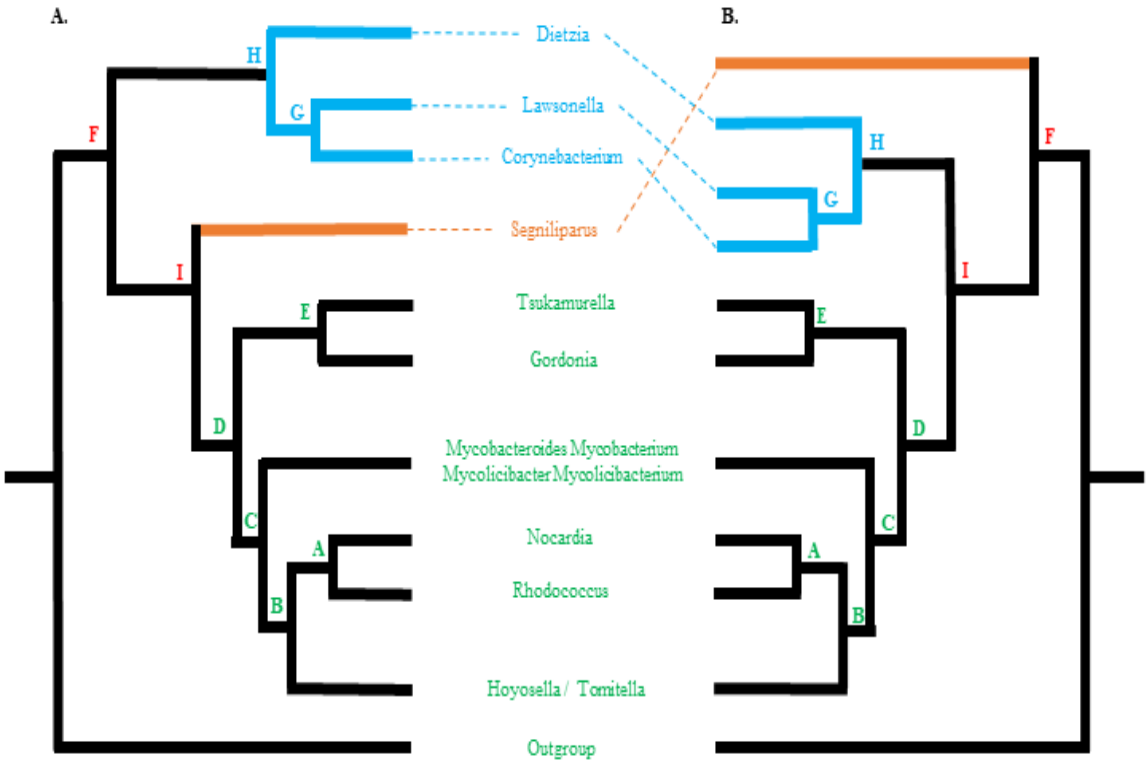


Figure 9*:* Topologies comparison between KeepOne and full tree of the 252 Actinobacteria. **(A)** Topology of the full tree before the KeepOne permutations were applied. **(B)** Topology of the KeepOne permutations (27 in total). Each letter (A – I) represents a node of interest in this analysis. Nodes I and F were the only ones to change from the full tree to any of the permutated trees. Like colored branches represent identical clades. Green nodes and species represent identical nodes / clades when comparing the full tree to each of the permutations topology.

With these parameters, for the 103 Actinobacteria subset, we identified 14 orthologs for a concatenated alignment length of 5,936 sites (after removing 3,513 sites with ≥ 25% gaps). For the 252 Actinobacteria subset we identified 142 orthologs for a concatenated alignment length 36,573 sites (after removing 23,208 sites with ≥25% gaps). Using both of these alignments independently, we reconstructed the phylogeny with RAxML with the estimated best model LG+G and 100 bootstraps. Each of the obtained trees, within subset, were compared to the Full tree that included all 103 species and 252 species, respectively, and to previously published reconstructions of the phylum Actinobacteria [116, 117].

To determine if taxon sampling had an effect on the estimated topology of the 103 dataset, we focused on the nodes that in Nouioui *et al.* [116] include *Nocardia*, *Rhodococcus*, *Amycolicicoccus* (recently renamed to *Hoyosella*), *Mycobacterium*, *Segniliparus*, *Gordonia*, and *Tsukamurella* (Fig. 8A; Nodes A – E) [116, 117]. We find that in all sampling scenarios the sister clade relationship between Nocardia / *Rhodococcus* and *Hoyosella* (*Amycolicicoccus*) remains stable (Fig. 8A; Node B), although bootstrap values are consistently higher when a single species is used (Fig. 8B). This suggests that the interaction of *Nocardia* and *Rhodococcus* during the tree reconstruction process may decrease the confidence in their association with *Hoyosella* (*Amycolicicoccus*). Next we conducted a similar analysis on the topologies of the 252 species data set. We also find that the sister clade relationship between *Nocardia*/*Rhodococcus* and *Hoyosella* (*Amycolicicoccus*) remains stable (Table 1; Node B) at 100 bootstraps for the full tree as well all but one of the permutations. In the full tree, *Skermania* grouped with the *Nocardia*. However, when it was a part of a keep one

scenario and all of the *Nocardia* were removed and it was used as the representative species for that group, the topology remained constant, but the bootstrap values decreased to 90. For this clade, as more *Nocardia*, *Rhodococcus*, and *Hoyosella* (*Amycolicicoccus*) were added, the topology remained constant suggesting that there were too few species to accurately reconstruct the clade in the 103 Actinobacteria analysis.

More interestingly, with the 103 Actinobacteria dataset we see changes in the deeper nodes depending on the species chosen to represent the *Nocardia/Rhodococcus* clade. All of the Nocardia species (*N. brasiliensis, N. cyriacigeorgica, and N. nova*) produce similar clustering patterns to that of the full tree with the exception of the Mycobacteria species that groups more closely with the *Gordonia* / *Tsukamurella* / *Segniliparus* clade (Fig. 8B; node C). The *Mycobacteriaceae* group (Myco* species) are composed of *Mycobacteroides*, Mycobacterium, *Mycolicibacter*, and *Mycolicibacterium*. When only *R. opacus* is included, we see very similar clusters to those of the full tree, but *Segniliparus* flips to the outside of the whole group (Fig. 8B). As for the 252 Actinobacteria dataset, we observe the Segniliparus node nesting just outside of the *Tsukamurella* / *Gordonia* / Myco* species / *Nocardia* / *Rhodococcus* / *Hoyosella* (*Amycolicicoccus*) / *Tomitella* clade for the full tree (Fig. 9A). However, when any permutation was implemented, *Segniliparus* flipped out to the ingroup root node and the clade containing *Corynebacterium* / *Lawsonella* / *Dietzia* flipped inward (Fig. 9B). Interestingly, there were no topological differences between the use of Nocardia or *Rhodococcus* species in the permutations, unlike what was seen in the 103 Actinobacteria analysis.

Table 1. Bootstrap values for all of the KeepOne scenarios and the standard (full) tree of 252 species. Node labels are based on the standard tree topology. Node F values were all 100 and node I were 99 in each of the permutated trees

	Bootstrap Values								
KeepOne Species	**Node A**	**Node B**	**Node C**	**Node D**	**Node E**	**Node F**	**Node G**	**Node H**	**Node I**
Nocardia asteroides	100	100	60	71	100	NA	100	100	NA
Nocardia brasiliensis	100	100	81	85	100	NA	100	100	NA
Nocardia cyriacigeorgica	100	100	62	76	100	NA	100	100	NA
Nocardia farcinica	100	100	62	77	100	NA	100	100	NA
Nocardia gipuzkoensis	100	100	80	86	100	NA	100	100	NA
Nocardia huaxiensis	100	100	54	79	100	NA	100	100	NA
Nocardia iowensis	100	100	82	86	100	NA	100	100	NA
Nocardia mangyaensis	100	100	59	70	100	NA	100	100	NA
Nocardia otitidiscaviarum GCF_007362295	100	100	59	81	100	NA	100	100	NA
Nocardia otitidiscaviarum GCF_014217765	100	100	42	72	100	NA	100	100	NA
Nocardia seriolae GCF_002356035	100	100	46	77	100	NA	100	100	NA
Nocardia seriolae GCF_018362975	100	100	44	78	100	NA	100	100	NA
Nocardia terpenica	100	100	67	82	100	NA	100	100	NA
Nocardia wallacei	100	100	71	80	100	NA	100	100	NA
Nocardia yamanashiensis	100	100	57	78	100	NA	100	100	NA
Rhodococcus aetherivorans	100	100	58	71	100	NA	100	100	NA
Rhodococcus coprophilus	100	100	72	78	100	NA	100	100	NA
Rhodococcus erythropolis GCF_000696675	100	100	67	78	100	NA	100	100	NA
Rhodococcus erythropolis GCF_015099595	100	100	67	78	100	NA	100	100	NA
Rhodococcus fascians	100	100	73	73	100	NA	100	100	NA
Rhodococcus globerulus	100	100	69	80	100	NA	100	100	NA
Rhodococcus hoagii	100	100	82	84	100	NA	100	100	NA
Rhodococcus koreensis	100	100	82	83	100	NA	100	100	NA
Rhodococcus opacus	100	100	82	83	100	NA	100	100	NA
Rhodococcus pyridinivorans GCF_000511305	100	100	63	75	100	NA	100	100	NA
Rhodococcus pyridinivorans GCF_016804345	100	100	62	76	100	NA	100	100	NA
Skermania piniformis	90	90	75	85	100	NA	100	100	NA
Full Tree	**100**	**100**	**100**	**99**	**100**	**78**	**100**	**100**	**100**

The discrepancies we observed with our taxon sampling results are also found in two other Actinobacteria phylogenetic studies that are not in agreement with each other nor with our study. The first study by Nouioui *et al.* [116], generated a phylogenetic topology that used as many species as possible (1,142 species) whereas the second study by Salam *et al.* [117] generated a phylogenetic topology based on 19 universal marker gene sequences of 404 type-strains. Both of these approaches, as well as ours, utilize concatenations but differ significantly on the initial amount of data used.

When comparing both studies to our full tree topology (no permutations) from the 103 Actinobacteria dataset, we note that the trees are substantially different with no overlapping nodes with Nouioui et al. [116], and only one node in common with the Salam et al. [117] (Fig. 8A; node A). However, when comparing these two studies to our 252 full tree topology (no permutations) we see a few more similarities. In the Nouioui *et al.* [116], we see that *Skermania piniformis* clusters with the *Nocardia* species and that node B (Fig. 9A) also exists. However, Node I, one of the backbone nodes, is completely different than what is depicted in the Nouioui et al. [116] tree. As for the Salam *et al.* [117] study, we see the similar clustering of Corynebacterium and Dietzia as well as the *Tsukamurella* and *Gordonia* (Fig. 9A). However, one interesting observation is that *Segniliparus* diverges at the base of the ingroup node in both Salam *et al.* [117] and our permutated trees. While the discrepancies between Nouioui and Salam topologies could have been attributed to differences in dataset and methods, our analysis shows that, even when the approach is the same, differences are identified, suggesting that taxon sampling is a primary cause of discrepancies among large phylogenies.

Interestingly, the discrepancies among our permutated trees are in nodes that are not always direct ancestors of the permutated clade. We also note that even when the topology remains stable, statistical support (bootstrap) for various nodes can change. These results highlight two important points: first, the effects of taxon sampling can be observed in nodes that are not direct ancestors of the species being analyzed; second, they can lead to a deeper understanding of nodes with low bootstrap values. In our 103 Actinobacteria example, nodes C and D in the full tree have bootstrap supports < 50 (Fig. 8B) but when species are sampled, node C in most cases is not recovered while node D is present in most permutations. This result would suggest that, despite the low bootstrap, node D subtends a stable clade while node C does not. In the 252 Actinobacteria example, the topology remains mostly stable but the bootstrap supports for nodes C and D, which are very high in the full tree, drop for all of the permutations.

2.10: Conclusion

Phylogenetics is an ever-evolving field that seeks to structurally organize species based upon genomic and proteomic data on the foundation of current taxonomical classifications. However, these taxonomical classifications are also ever-changing as we discover and sequence new species and compare them to previously classified species. Phylogenetics allows us to hypothesize evolutionary connections between differing species, while in-turn, allows us to test for proper taxonomical classifications of these species. For example, in the Nouioui *et at.* [116] tree, there is a *Rhodococcus* species (*R. kunmingensis*) grouping with the *Nocardia* species and *Skermania piniformis*. When this tree was generated, this is where *R. kunmingensis* fell based upon comparative orthology

of 16s rRNA. However, due to its clustering outside of the core *Rhodococcus* clade, they suggested its reclassification as a new *Nocardia* species. Eventually, it was reclassified into a whole new genus, *Aldersonia*. Orthology and tree reconstructions play a crucial role in testing taxonomical classifications even though species need to be classified before they can be a part of a tree reconstruction. This delicate dance between taxonomical classification of a new species and testing of its evolutionary place within a phylogeny is nuanced by human biases and computational constraints. That is why there is a large importance on attempting to optimize techniques used in these types of analyses to increase phylogenetic accuracies and limit these human biases.

One thing that is guaranteed in the field of phylogenetics is that there will always be more data to classify and structurally organize. Because of this, there is a two-pronged issue. First, as data flows in, computational constraints become more and more of a burden on large-scale analyses attempting to stabilize trees that contain large numbers of species and large numbers of taxa. For example, as more species representing the same genus are added to an analysis we tend to see more consistent structural organizations, as seen by the higher bootstrap values of the *Rhodococcus* / *Nocardia* / *Hoyosella* (*Amycolicicoccus*) and *Gordonia*/*Tsukamurella* clades from the 103 Actinobacteria species analysis (Fig. 8B) to the 252 Actinobacteria species analysis (Table 1). The increase of bootstrap values of shallow nodes is also important as it can lead to the identification of misclassified species (as seen in Nouioui *et al.* [116] and, potentially, a *Corynebacterium* species (*Corynebacterium cyclohexanicum*) in our 252 species analysis). Unfortunately, computational constraints become more apparent as we move up the taxonomical classification scale (i.e., speices → genus → family → order, etc.),

which can limit the number of species that can be added to stabilize the phylogenetic tree. The second issue is the attempt to minimize data to represent large clades from few user-selected species. As more species are added to an analysis, an exponential increase in computational run-time is observed for large-scale analyses. Utilizing a smaller subset of species to represent a specific clade in a larger tree is an inevitability but its effects are rarely tested. In some cases, there will be little to no effect. For example, in our 103 Actinobacteria species analysis, the *Gordonia*/*Tsukamurella* clade is always reconstructed with bootstraps of the keep one scenario and full tree ranging from 86 – 94 (Fig. 8) and, similarly, the bootstraps of the *Gordonia* / *Tsukamrella* in the 252 Actinobacteria species analysis are all 100 (Table 1). This also holds true also when examining the keep one scenario for the 252 Actinobacteria species analysis where we see that the *Nocardia*/*Rhodococcus* node (Fig. 9; node A) is consistent across all of the keep one scenario runs with bootstraps of 100 (Table1). However, in other cases, topology and/or bootstrap values can change. In the 252 Actinobacteria dataset, we see a flip in topology between the full tree and all of the KeepOne permutations. We see Segniliparus nested within the *Dietzia* / *Lawsonella* / *Corynebacterium* clade for the full tree (Fig. 9; node I) but in all of the permutated trees, *Segniiparus* is the sister lineage of all species considered in this analysis (Fig. 9; node F). Thus, the testing of how a single species affects the topology of a phylogeny is extremely important when that species is being used to represent a larger clade.

In conclusion, taxon sampling is an important methodology for phylogenetic reconstructions and can be easily deployed with the use of the PATS pipeline. It can be used in heavily mall tree reconstructions that have high levels of species representation

per each genus to test for paralogy or species misclassification. More importantly, it can be applied for the testing of representative species in large-scale tree reconstructions. Implementation of a keep one or a remove one scenario easily allows for the generation of a comparative framework to see which species might or might not be the best to represent the larger group. PATS is even useful for simple tree creation where taxon sampling is either not needed or already tested. The fully automated nature of PATS lends to its ease of use and simplicity of deployment.

CHAPTER 3

TESTING PRIMARY *VS* SECONDARY CALIBRATIONS IN A SIMULATED ENVIRONMENT FOR MOLECULAR CLOCK APPLICATIONS

3.1: Introduction

Topology is a key outcome of phylogenetic reconstructions, but other information can be obtained with further analyses. For example, the molecular clock approach leads to the estimation of time in phylogenetic reconstructions. Simplistically, this can be done by accounting for nucleotide evolutionary rates and branch length distances via topologies. Since rate equals genetic distance over time, this can be used to apply time to a phylogenetic reconstruction. This is often done by adding calibration points, such as fossil information, that is dated (primary calibration point). However, as more and more molecular clock analyses are run, consistent divergence times for some nodes that lack fossil information can be used as a secondary calibration point for other analyses. This approach has been criticized as it may propagate errors from one analysis to the next.

The use of calibrations to estimate absolute times in a phylogeny is a source of controversy for many reasons. Among these are that (i) few are available from independent sources (e.g., fossil record), (ii) their phylogenetic placement can be incorrect, especially in cases of uncertain fossil identification or topological location within a phylogeny, and that (iii) calibration constraints (and the internal distributions between them) are heavily debated, although new methods to estimate probability

densities of node ages are being developed [119, 120, 129–132, 121–128]. Despite these issues, molecular clock analyses cannot avoid using calibrations if absolute time estimates are the ultimate goal. Alternative approaches that have been explored include the direct use of substitution rates to estimate divergence times, but these present other challenges such as their applicability when, in many cases, these rates are measured from laboratory-grown strains and may not accurately represent rates of strains in the environment [126, 133]. A recently developed new strategy, instead, suggests the use of horizontal gene transfer events as time calibrated constraints in a phylogeny [134, 135]. While this method holds promise, it currently has not been applied widely and the reconstruction of horizontal gene transfer events useful as calibration constraints is a challenging endeavor as it relies on clear gene exchange patterns between groups, one of which should also have primary calibrations. Thus, challenges for these alternative calibration strategies are equally difficult to overcome and, for now, limited improvements have been made. Therefore, the process of assigning time constraints to some nodes in a phylogeny still remains the primary source of information to obtain absolute node times, despite the unbalance between the amount of information (i.e., number of nodes that can be calibrated) available and what would be desirable to obtain accurate time estimates. This unbalance between availability and need has led researchers to test a variety of alternatives to meet the new needs by increasing calibration sources. These new strategies are especially important in very large phylogenies due to the increase in the ratio of unknown to calibrated nodes and, thus, the potential increase in error propagation of estimates for nodes that are far from a calibration point often caused by rate variation among branches [126, 136–139].

Some examples of alternative strategies to improve the number and quality of available calibrations include (i) the definition of boundaries either based on the time boundaries of the geologic stratum in which the fossil is found or based on an accurate phylogenetic placement of extinct taxa [119, 124, 140, 141]; (ii) the selection of representative taxa, which effectively decreases the unknown/calibrated node ratio [136, 138]; and (iii) the use of secondary calibrations [123, 126, 142]. In this aim we focus on secondary calibrations, which are molecular time estimates obtained from previous molecular clock analyses that were calibrated using independent evidence (primary calibrations). The strongest advantage of using derived (e.g., secondary, tertiary) calibrations is that it effectively provides an infinite source of calibrations, only constrained by the number of steps (nodes) a researcher chooses between calibrated and unknown nodes. However, this advantage is dependent on one fundamental question: does the use of derived calibrations result in accurate timetree estimates? In other words, does the inevitable error associated with molecular time estimate propagates to and increases in other nodes if these estimates are used as calibrations?

Two past studies have addressed this question with simulation and empirical data analyses, using both Bayesian and Maximum-likelihood based methods [123, 142]. Both found similar outcomes, with secondary calibration estimates being younger than expected and with overly narrow confidence intervals leading to small uncertainties around inaccurate estimates. These results supported previous evidence against the use of secondary calibrations reinforcing the practice of avoiding them, if possible (e.g., [126, 137, 143, 144]). However, questions about the performance of secondary calibrations remain. For example, how is the error from primary calibrations compounded into

estimates based on secondary calibrations? Can the performance of secondary calibrations be predicted based on uncertainties in the primary calibrations? Is the performance of secondary calibrations worse than that of few primary calibrations that are phylogenetically distant from the nodes of interest (distant primary calibrations)?

To address these questions, we designed a simple simulated scenario with the aim of testing if there are predictable patterns when secondary calibrations are used. For this purpose, we used RelTime to estimate times because of its minimal assumption requirements and speed of analyses [145, 146]. In our scenario we used two nested trees that share one overlapping ingroup node that is used as secondary calibration. In one of the two trees we also selected three nodes to act as primary calibrations. These were used with increasingly larger uncertainties in their boundaries (from 0 to 20% departure from the true, simulated time) and biases (either balanced around the true time or skewed towards younger or older times). Based on the results from previous studies, we expected that estimates based on secondary calibrations would be consistently and precisely underestimated relative to the use of primary calibrations and the true times. Instead, we found that secondary time estimates are generally overestimated by approximately 10% but with low precision (large confidence intervals) and with overall patterns that are clearly predictable. These results suggest that our understanding of the accuracy of secondary calibrations is still incomplete and more comprehensive testing is required to determine their effect if used in empirical datasets.

3.2: Methodology

3.2.1: Dataset We started from a main tree of 248 species represented in a tree of life [147]. This main tree was split into two subtrees (Fig. 10), tree A (173 species) and tree B (71 species), that represent two clades and maximize the size of the dataset in each tree.

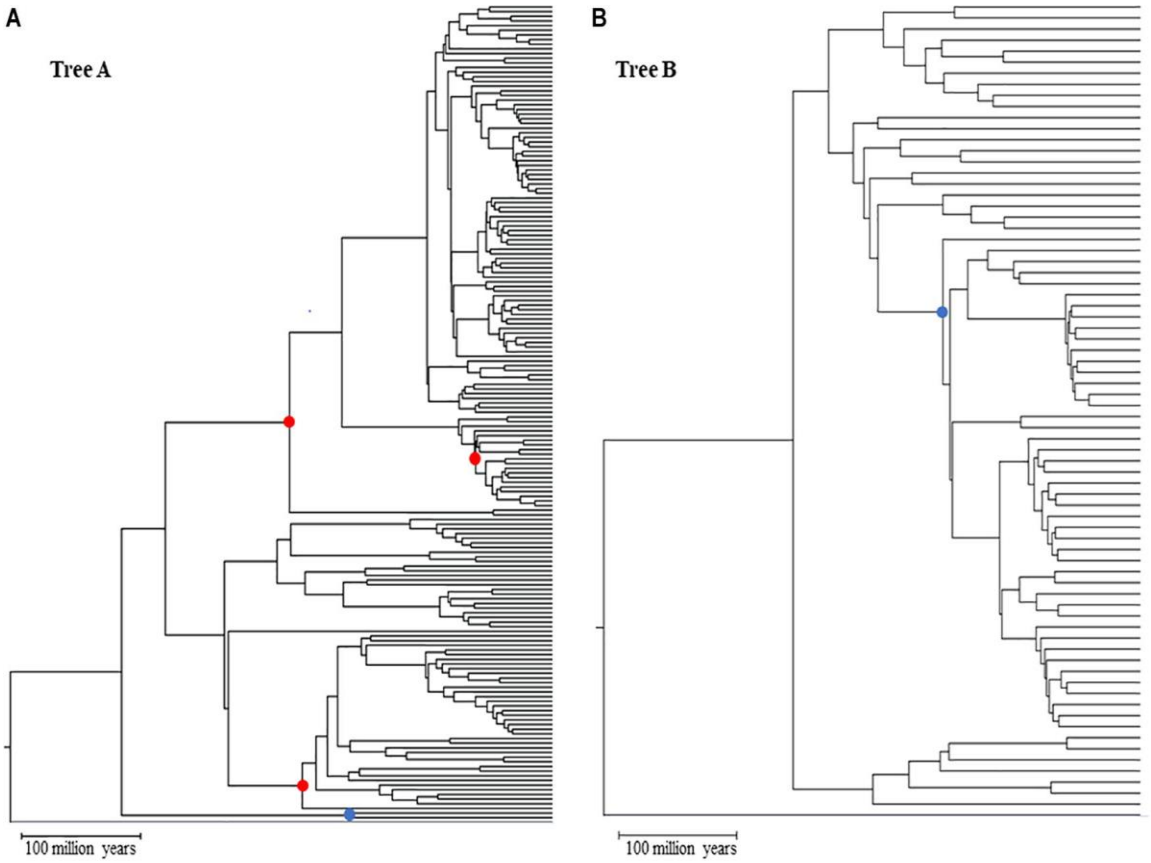


Figure 10: Topologies used in simulated analyses for Tree A **(A)** and Tree B **(B).** Red dots: primary calibration nodes; blue dot: overlapping node between Tree A and Tree B used as secondary calibration in tree **(B)**. Gray branch: outgroup.

In selecting both tree A and tree B we ensured that 2 lineages present in tree B would remain also in tree A. These two lineages were arbitrarily chosen. This setup created two nested phylogenies that were used to test hypotheses on calibrations'

performance. To simulate multiple genes, we used a set of 446 empirical parameters (e.g., length, GC content, initial evolutionary rate) and altered the main timetree according to an autocorrelated model ($\nu = 1$) that resulted in estimated rates of up to ± 25% of the mean rate (Fig. 11) [148, 149]. This effectively created 446 phylogenies with different branch lengths but same topology. These parameters were given to SeqGen to simulate genes under an Hasegawa-Kishino-Yano (HKY) model [150]. Ten random sets of individual genes were then concatenated to reach a length of at least 30,000 sites (30,029 – 30,725). In addition, we also created one concatenation with all genes (603,999 sites) and two concatenations of half the number of genes (223 genes per concatenation) with lengths of 273,812 and 330,187. Each of these concatenations were used independently in downstream analyses. Patterns between the 30k, half, and full concatenations were similar. Therefore, we discuss results from the 30k concatenations because they allow us to evaluate the variance of estimates among datasets. For primary calibrations, three nodes from tree A were chosen: a relatively shallow node at 63.9 million years ago (MYA), and two that were deeper in the tree but in two different clades (209.4 MYA and 220.2 MYA). The overlapping node between tree A and B has an intermediate depth (167 MYA) within tree A and is centrally placed within the topology of tree B (Fig. 10). These primary and secondary calibrations were chosen to minimize the effect that biased location (e.g., all within one clade) and divergence times (e.g., all young nodes) may have on the accuracy of estimations.

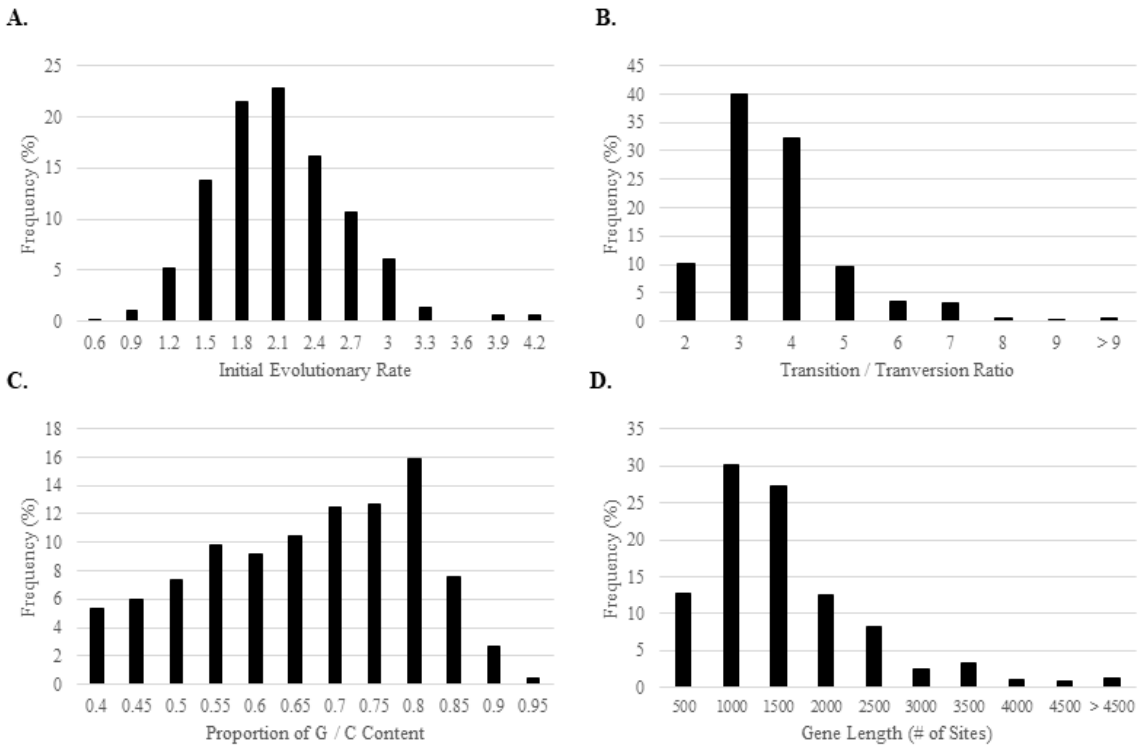


Figure 11. Empirically derived parameters used to simulate alignments A: initial evolutionary rates; B: transversion and transition ratios; C: proportion of guanine and cytosine bases; and D: alignment length of individual genes.

3.2.2: Time estimation Time estimates for each concatenation were calculated using RelTime as implemented in the command-line version of MEGA X [99]. Each analysis was run on the Michigan State University HPCC-iCer cluster using HKY model, uniform rates among sites, all sites, a maximum likelihood estimator, and local clocks. Our goal was to explore the accuracy of time estimates for the nodes in tree B when different types of calibration were used. Thus, we used 3 combinations of calibrations, all with minimum-maximum constraints: tree A with three primary calibrations (Fig. 10, red dots) plus tree B with one secondary calibration (this is the overlapping node between tree A and B; Fig. 8, blue dot) [B_secondary]; tree B with one primary calibration (same node as the previous combination) (Fig. 10, blue dot) [B_primary]; tree AB (the

combination of trees A and B) with the same three primary calibrations used in tree A (Fig. 10, red dots) but that, effectively, are distant from the nodes in tree B [B_distant_primary]. Additionally, each of these combinations was tested for 7 different scenarios that were meant to account for increasing uncertainty on primary calibrations: three of these scenarios had increasingly larger uncertainties, from 0 to 20%, but spread evenly around the true time (0 balanced (0B), 10 balanced (10B), and 20 balanced (20B)); two scenarios had the error skewed towards younger times (10 low (10L) and 20 low (20L)) and two with the error skewed towards older times (10 high (10H) and 20 high (20H)). This set up allowed us to test if and how the error in primary calibrations propagates to estimates based on secondary calibrations.

3.2.3: Measures of accuracy We assessed the outcomes of our molecular clock analyses with a number of measures. First, we measured the percent departure of the estimated times (ET) from the known (simulated) true times (ET accuracy). Second, we calculated the frequency of confidence intervals (CIs) that included the true times (TT) (CI accuracy). Third, we calculated the range of each CI normalized to the depth of the node ((maximum – minimum boundary)/TT) (CI precision). Fourth, we measured the distance of each CI boundary to the true time (|(CI boundary – TT)/TT)|) (CI skewness). We applied the latter measure only to the overlapping node between tree A and B to evaluate the effect of skewness on estimates based on secondary calibrations. Each of these measures was applied to all analyses. We report the averages across the 10 concatenations with ± 1 standard deviation in parenthesis because no significant difference was detected among individual concatenations.

3.3: Results

Generally, secondary calibrations are used when primary calibrations are not available or are believed to be too few to provide accurate estimates on nodes that are distantly related. While many studies have used this approach as a last-resort method, a systematic evaluation of their performance might open up the use of secondary calibration to more (and larger) phylogenies, thus expanding the applicability of molecular clocks to complex datasets. To investigate and quantify the amount of error associated with secondary calibration, we used a simulated approach by creating two nested trees (A and B) in which one node estimate from primary calibrations in A is used as a secondary calibration in B. We then quantified the accuracy of estimated times (ET) *vs.* simulated true times (TT) and the properties of the confidence intervals (CIs) in both trees relative to the type of calibration used. Each analysis was repeated for a series of scenarios with variable levels of uncertainty in the primary calibrations to investigate how estimates derived from secondary calibrations may be affected (see Methods).

In practice, we simulated hundreds of nucleotide alignments with a variety of empirically derived parameters (length, initial evolutionary rate, transition/transversion ratio, GC content) for a phylogeny of 248 lineages that was split into two nested trees (A with 176 species and B with 74 species). We used variable numbers of genes in concatenation to obtain the alignments used to estimate divergence times. On three nodes in A, we applied primary calibrations with varying levels of uncertainty around their true time (Table 2). Then, the CI of the molecular time estimate for the common node between A and B was used as secondary calibration for tree B.

Table 2. Scenarios of varying uncertainty around the true time (TT) for primary calibration boundaries. Each scenario was applied to all three primary calibrations at once and run independently from the other scenarios. The uncertainty associated with the boundaries varies from a minimum of 0% (± 1 million year of the true time) to a maximum of 20%. my: million years.

Scenario	Calibration Boundaries	
	Minimum Time	**Maximum Time**
0B	-1 my	+1 my
10B	-5% of TT	+5% of TT
20B	-10% of TT	+10% of TT
10L	-10% of TT	+1 my
10H	-1 my	+10% of TT
20L	-20% of TT	+1 my
20H	-1 my	+20% of TT

Our evaluations of time estimates' accuracy rely on two basic measures: similarity of the estimated time to the simulated true time (ET accuracy) and the general properties of the CIs (accuracy, precision, skewness). In addition, we also compared estimated times from secondary calibrations to those obtained from B_distant_primary calibrations when the two trees were combined.

3.3.1: Assessing accuracy of primary calibrations The first step was to measure the trends in accuracy of estimated times and CIs from primary calibrations in tree A. On average, the ET differed from TT by approximately 3% with a tendency to underestimate the results (average of slopes = 0.97, stdev = 0.06) which reflects a

generally higher confidence on younger (minimum) calibration boundaries than older (maximum) ones (Fig. 12; black triangles, Table 3, and Appendix Figure A1). The observed trends in accuracy were as expected relative to the error associated with calibration boundaries: balanced calibrations (0B, 10B, and 20B) performed the best, boundaries skewed towards younger times (10L and 20L) produced underestimated times and boundaries skewed towards older times (10H and 20H) produced overestimated times.

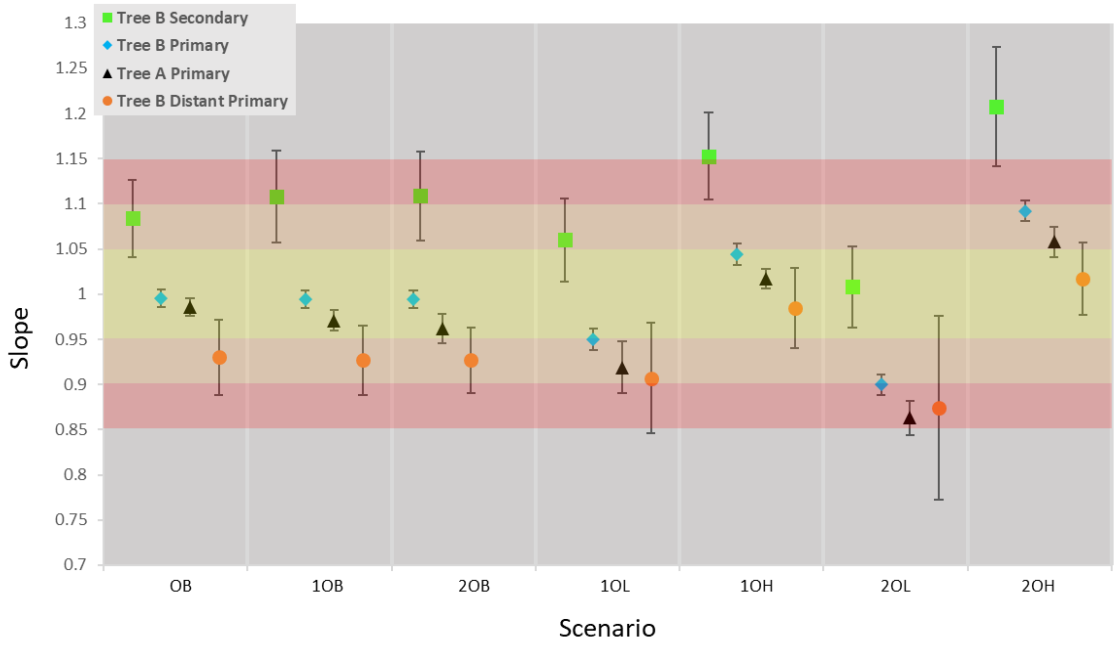


Figure 12: Comparison of average slopes for each of the seven scenarios in the four-calibration setting. Gradient of color represents the accuracy of the estimates yellow is 5%, orange 10%, red = 15%, grey > 15%. Each data point represents the average slope of the 10 concatenations of true time *vs.* estimate time with 1 standard deviation.

Table 3. Average slope of estimated times *vs.* simulated true times (ET accuracy) for ten 30K concatenations. The slope was calculated from a best fit linear model with intercept = 0. Values for 1 standard deviation are shown in parentheses. (See Appendix Figure 1 – Appendix Figure 5 for graphical representation).

Scenario	ET Accuracy			
	Tree A Primary	**Tree B Secondary**	**Tree B Distant Primary**	**Tree B Primary**
0B	0.99 (0.0097)	1.08 (0.0427)	0.93 (0.0414)	0.99 (0.0097)
10B	0.97 (0.0110)	1.11 (0.0507)	0.93 (0.0386)	0.99 (0.0097)
20B	0.96 (0.0162)	1.11 (0.0493)	0.93 (0.0365)	0.99 (0.0097)
10L	0.92 (0.0288)	1.06 (0.0457)	0.91 (0.0611)	0.95 (0.0115)
10H	1.02 (0.0106)	1.15 (0.0485)	0.98 (0.0443)	1.04 (0.0117)
20L	0.86 (0.0189)	1.01 (0.0449)	0.87 (0.1019)	0.90 (0.0115)
20H	1.06 (0.0169)	1.21 (0.0660)	1.02 (0.0397)	1.09 (0.0113)
Average	*0.97 (0.06363)*	*1.10 (0.0643)*	*0.94 (0.0480)*	*1.00 (0.0617)*

On average, the CI ranges included the true time in 87% of the cases with the 0B scenario being the most likely to fail (78% of CIs include the TT) (Table 4). This is expected as narrower calibration boundaries generate narrower CI ranges that are less likely to include the TT in light of the underestimated divergence times. Similarly, in an average of 87% of all the cases the CI included the true time for the node overlapping in trees A and B and, in those cases in which the TT was not included, the minimum boundary was < 5% higher than the TT (the maximum boundary was never younger than the TT). In only one scenario, 20H, the difference between TT and minimum boundary was 6.5%, which is expected given the large uncertainty skewed toward deeper times. Moreover, in 93% of the cases the CI of the overlapping node was skewed toward older times which means that, when used as a secondary calibration, this node is expected to behave similarly to the "high" simulated scenarios (see below). Overall, the trends observed from different calibration scenarios in tree A are as expected and provide an accurate basis for the estimation of times using a secondary calibration.

3.3.2: Assessing accuracy of secondary calibrations……..Using the CI estimated for the overlapping node between tree A and tree B, we obtained secondary divergence times and measured the accuracy of these secondary node ages relative to true times and to estimated times from primary calibrations on distant nodes (we refer to this scenario as B_distant_primary). The first measure is unrealistic in real cases but allows us to quantify the overall error produced by secondary calibrations relative to the true times, while the second measure leads to a quantification of the error introduced by secondary calibrations compared to distant primary ones.

Table 4. Confidence intervals (CIs) accuracy (proportion of CIs that include the simulated true time) for each of the seven scenarios. Values shown are averages of all 30k concatenations with 1 standard deviation in parenthesis.

Scenario	CI Accuracy			
	Tree A Primary	**Tree B Secondary**	**Tree B Distant Primary**	**Tree B Primary**
0B	0.78 (0.1495)	1.0 (0.0044)	0.98 (0.0222)	0.91 (0.0287)
10B	0.87 (0.1333)	1.0 (0.0044)	0.97 (0.0306)	1.0 (0.0048)
20B	0.95 (0.0704)	1.0 (0.0044)	0.98 (0.0181)	1.00 (0.0000)
10L	0.86 (0.0608)	1.0 (0.000)	0.89 (0.1139)	0.99 (0.0000)
10H	0.84 (0.1590)	1.0 (0.0059)	0.99 (0.0097)	0.99 (0.0125)
20L	0.88 (0.1223)	1.00 (0.0000)	0.85 (0.1528)	1.0 (0.0042)
20H	0.89 (0.1658)	0.99 (0.0240)	1.0 (0.0014)	0.99 (0.0097)
Average	*0.87 (0.0525)*	*1.0 (0.0044)*	*0.95 (0.0589)*	*0.98 (0.0313)*

Contrary to previous studies, our results show an average 10% (± 6%) overestimation of molecular time estimates against true times with the strongest overestimation in the scenarios with the largest inaccuracy on the maximum boundary (10H, 20H) (Fig. 12, green squares; Table 3, and Appendix Figure A2). This is predicted from the CI boundaries of the secondary calibrations that are most strongly skewed towards older times in the "high" scenarios. Interestingly, the amount of inaccuracy produced by secondary calibrations is comparable to the average 7% (± 6%) departure of the estimates from the true times produced by distant primary calibrations, although in these cases the node ages are generally underestimated (Fig. 12, orange circles; Table 3, and Appendix Figures A3 and A4).

Despite the overestimation produced by the secondary calibrations, > 99% of CIs include the true time (Table 4). The high probability of the true time being included in the CI of each node is due to the large CI range estimates (78 – 95% of the true time) (Table 5). This is approximately double the size of the CIs obtained from distant primary calibrations (42 – 53% of the true time) and from primary calibrations. The larger size of the CIs when secondary calibrations are used reflects the larger uncertainty in the calibrating range. Indeed, while the primary calibrations were allowed to have at most a 20% uncertainty, the CIs of the node used as secondary calibration have, on average, double that amount, thus producing two-fold larger CIs in the estimated nodes.

It is possible that the results obtained in the estimation of node ages for tree B with secondary calibrations could have been driven by anomalies in branch lengths and evolutionary rates specific to this phylogeny rather than the type of calibration used. To

identify these potentially confounding factors we first applied a primary calibration on the same node that was used as secondary. If the location and branch lengths associated with the calibration node were biasing the results, the accuracy of estimated times should have been lower even with primary calibrations. Instead, we observed similar accuracy and precision to what was obtained with three calibrations in tree A (Fig. 12, blue diamonds; Appendix Figure A5, and Table 3; ET on average within 5% of TT, > 98% of CIs include the TT (Table 4), and the CI precision is approximately 40% (Table 5)).

Then, we compared the relative times (before calibrations are applied) of nodes represented only in tree B, only in tree A, and in the combined tree AB. If trees A and B differed substantially in the relative rates of their branches, we would expect that the relative times would not be comparable to those obtained with the combined tree AB. Again, we found the opposite result, which suggests that evolutionary rates do not differ significantly between trees (Fig. 13). These results show that the estimated times obtained are driven primarily by the choice of calibration and that, therefore, they are a valid measure of calibration performance.

These results show predictable trends for the estimates of secondary calibrations that closely mirror the uncertainty of the primary calibrations. Additionally, they show that, at least in this simulated scenario, the absolute accuracy of using a secondary calibration is similar to that of using distant primary calibrations, although the precision is approximately half.

Table 5. Confidence interval (CI) precision relative to the simulated true time. Averages of the 30k concatenations with 1 standard deviation are shown in parentheses.

Scenario	CI Precision			
	Tree A Primary	**Tree B Secondary**	**Tree B Distant Primary**	**Tree B Primary**
0B	0.29 (0.0507)	0.78 (0.1409)	0.45 (0.1420)	0.27 (0.0548)
10B	0.32 (0.0606)	0.78 (0.1740)	0.43 (0.0826)	0.38 (0.0516)
20B	0.45 (0.0614)	0.89 (0.1756)	0.49 (0.0943)	0.51 (0.0474)
10L	0.32 (0.0543)	0.78 (0.1626)	0.42 (0.0777)	0.37 (0.0487)
10H	0.33 (0.0655)	0.82 (0.1867)	0.45 (0.0886)	0.40 (0.0540)
20L	0.44 (0.0567)	0.82 (0.1316)	0.48 (0.0827)	0.50 (0.0353)
20H	0.47 (0.0670)	0.95 (0.1780)	0.53 (0.1045)	0.54 (0.0525)
Average	*0.37 (0.0749)*	*0.83 (0.0065)*	*0.46 (0.0039)*	*0.42 (0.0955)*

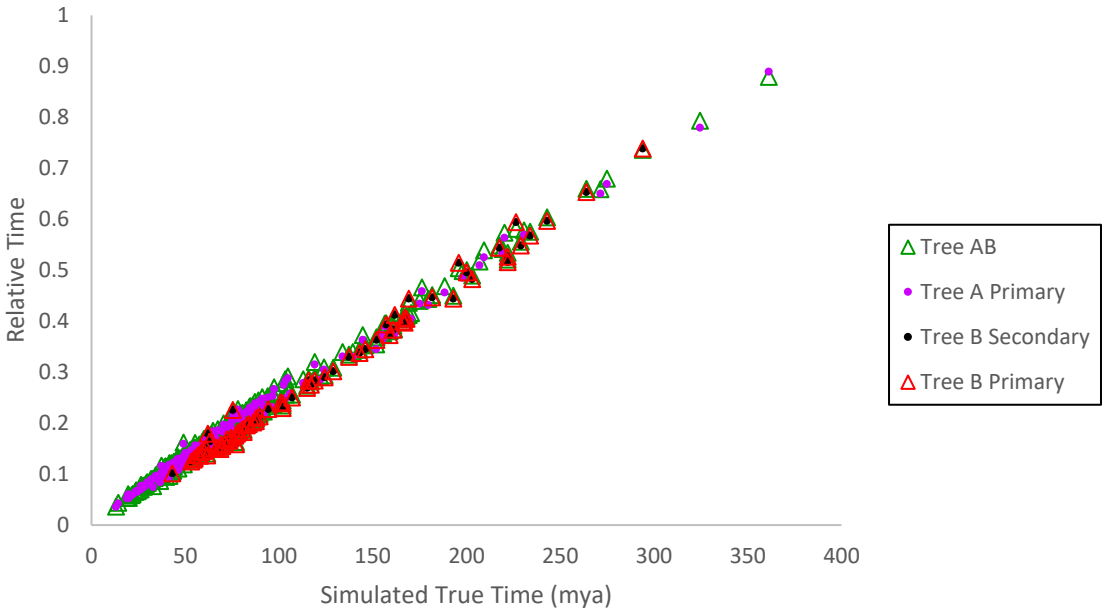


Figure 13: Simulated true times *vs.* relative times from RelTime before calibrations are applied. As expected, tree B relative times before primary or secondary calibrations are applied and are the same. Each of the four scenarios contain similar relative time outputs before being calibrated.

3.4: Conclusions

The measures we used to determine the effects of the use of secondary calibrations are the typical ones of molecular clock assessment studies: the similarity of true times and estimated times, the frequency of CIs that include the true time, and the precision of CI (their range relative to the age of the node). In addition to these, we also considered the relative error (TT vs ET) and CI precision of secondary calibrations *vs.* distant-primary calibrations and the predictability of secondary estimates based on the simulated scenario. Surprisingly, our results are opposite to those found by two recent studies: our results show that secondary estimates are, in general, overestimated (by approximately 10%) with poor precision (large CIs). While these results are far from optimal, they show that our understanding of secondary estimates is still incomplete and

that their dismissal might be premature. Perhaps more interestingly, we also found that the magnitude of the error in secondary estimates is approximately the same as that produced by the use of distant primary calibrations but in the opposite direction (secondary calibrations overestimate, distant-primary underestimate). This result is significant because one of the strategies commonly adopted to avoid using secondary calibrations is to increase the dataset size to obtain one or more primary calibrations[138]. These will inevitably be far away, in the phylogenetic sense, from the nodes of interest potentially leading to errors, as we see in our simulations. However, an advantage of using distant primary calibrations would be two-fold higher precision (narrower CI ranges) that does not come at the expense of a lower probability of including the true time. It should be noted that in our simulated scenario primary calibrations are given with minimum and maximum boundaries which are expected to increase accuracy and precision. It is possible that the precision of distant-primary calibrations would be negatively affected in empirical analyses that do not use maximum boundaries.

A deeper analysis of estimates from proximal primary, distant primary, and secondary calibrations can explain the trends observed. First, the large CI ranges from secondary calibrations are caused by the large uncertainty in the boundaries derived from the primary calibrations. Indeed, these boundaries include a 30 – 40% error, which is almost double the maximum amount of error assigned to primary calibrations. Therefore, these results suggest that estimates based on secondary calibrations incorporate the error present in the primary estimate in addition to their own. Second, the directionality of the secondary estimated errors (over- or underestimation) is also clearly dependent on the

skewness of the primary calibration boundaries. For example, in our simulated cases, we saw that CIs for the overlapping node were almost always skewed towards older times by approximately 30%, driving the observed overestimation. Thus, careful choice of accurate primary calibrations is key when secondary calibrations are to be used. Unfortunately, errors associated with primary calibrations are often unknown in empirical data (but see [99]) but knowing that these errors are included in the estimates based on secondary calibrations with predictable trends allows to test the plausibility of different evolutionary hypotheses based on what is known of the primary calibrations.

A question that remains open is why our results differ so strikingly from those of previous studies. While additional analyses will be necessary to provide an answer, a few hypotheses are possible: first, previous studies used primarily Bayesian methods that are known to depend strongly on priors [128, 139]. It is possible that the priors affected the results but the magnitude of this effect, if present, is unknown. Second, the two previous studies did not take into consideration the error associated with the primary calibrations. In one case [142] it was not simulated and in the other [123] it was not known because an empirical dataset was used. From the analyses of multiple scenarios (from many different studies) that the authors carried out it is obvious that the youngest estimates based on secondary calibrations are obtained when only young proximal primary calibrations are used, which is the same trend we observe in our analyses and in other, earlier, studies [99, 151].

Overall, our simulation study shows that our current understanding of the performance of secondary calibrations is still incomplete and that their dismissal from

implementation in favor of other solutions (e.g., distant primary) might not produce the desired increase in accuracy. The possibility of explaining the biases observed in secondary estimates based on the parameters specified in each scenario (i.e., uncertainty in primary calibrations) also shows that the performance of secondary calibrations can be predicted and, therefore, used as a testing tool for different evolutionary scenarios. Therefore, we suggest that rather than avoiding secondary calibrations, they are used and compared with distant primary ones to test the limits of the parameter space of plausible evolutionary scenarios of divergence times.

CHAPTER 4

TESTING THE EFFECT OF SPECIES CHOICE ON TIME TREE RECONSTRUCTIONS AND ITS IMPLEMENTATION INTO THE PATS PIPELINE

4.1: Introduction

When performing phylogenetic tree reconstructions, it is clear that the species chosen to reconstruct a tree (taxon sampling) will affect the final topology of that reconstruction. The species utilized in phylogenetic reconstructions can, on one extreme, be all of the species currently sequenced in a given taxonomic group or, on the other extreme, be chosen as representatives of a much larger group at the discretion of a researcher. This discretion innately adds biases that, if not assessed, could affect the final output. The inclusion of all species in each taxonomic group would alleviate these biases, however as dataset sizes grow in number, this approach becomes improbable due to computational and time constraints. Additionally, as sequencing efforts are not evenly applied to different phylogenetic groups, even using all available sequenced species will result in some groups represented by few lineages. The goal of a phylogenetic reconstruction is to generate a well-supported phylogeny which means that quantifying the discrepancies caused by different species choices is key. However, assessing the effect of species choice is a rarely performed step in most phylogenetic analyses because of constraints such as time, computational power, and data availability. With the multitude of currently published phylogenetic reconstructions lacking proper taxonomic sampling assessment, future interpretation of these non-tested taxon sampled

phylogenetic reconstructions could lead to cascading inaccuracies in scientific discoveries. A recent study performed by our group [21] shows that varying the species choice while maintaining identical parameters during phylogenetic reconstructions can lead to varying tree topologies, which can have undesirable effects on the interpretation of evolutionary histories. However, this study did not address the two other major components of a phylogeny: branch lengths and time, both fundamental to comprehensively reconstruct the evolution of life. Extending taxon sampling assessment into time tree reconstructions (molecular clock analyses) provides insight into whether or not there are additional effects that carry over into post-reconstruction analyses.

The main outputs from a phylogenetic tree reconstruction are the tree topology (e.g., mapping of species with respect to speciation events) and the branch lengths for each of the lineages within the topology. Branch lengths typically represent the number of nucleotide/amino acids substitutions per site, which is a representation of evolutionary distance between two nodes or lineages. In its simplest form, this evolutionary distance is calculated by taking a sequence alignment and dividing the number of nucleotide/amino acid differences by the total sequence length [46, 47]. However, branch lengths can also be represented in time units, either relative or absolute. RelTime [145, 146], a software built into the MEGA software suite [99, 152], takes in a tree topology and alignment for a set of species and estimates evolutionary times. Relative times are constrained by the topology and branch lengths used in the analysis, meaning they are relative to the root of the tree which, in itself, is not timed. Thus, relative times only allow for internal comparisons among nodes (e.g., node A is *older* than node B). However, calibrating the

relative time tree, typically using fossil records, allows for absolute time estimation (e.g., node A is 100 million years old, node B is 75 million years old). Adding in these timed calibrations into the relative tree allow for estimated timing of all of the nodes in the tree via a Bayesian (i.e., MCMCTree [92]) or a maximum likelihood approach (i.e., RelTime [145, 146]). As larger numbers of calibrations are used to time a phylogenetic tree, the accuracy of the estimated absolute time tree tends to increase [137].

The outputs from the PATS pipeline, specifically the topologies and branch lengths from the permutations and full tree, create an excellent comparative framework environment to test for these taxon sampling effects. For the purposes of the branch lengths and divergence time analyses, we used the 103 Actinobacteria dataset. This choice was dictated by the large length of the alignment in the 252 Actinobacteria dataset (almost 200,000 sites), which substantially increased the computational times even for the fastest method we used, RelTime. First, we compared each of the branch lengths for the nodes of interest (Fig. 8) across all of the 4 keep one scenario phylogenetic trees with each other and then with the full tree that contains all species. Second, we implemented secondary calibrations (two different calibrated nodes) into each topology, independently, to test for effects on absolute time calculations. We calibrated node C and node D (Fig. 8) because node C is the direct ancestor of the permutated clade and node D is more distant from the permutated clade. Because currently no primary calibrations are available for the *Actinobacteria* dataset we will be relying on secondary calibrations only to time the tree, giving us the opportunity to explore the effect of these kinds of calibrations in a real dataset (instead of a simulated dataset as done in Aim 2).We compared divergence times of shared nodes across each of the calibrated permutations with the full tree generating

linear models to determine differences across these taxon-sampled node estimates. Comparison of the full tree to any one of the permutations will either depict one of two statistical categories we have defined: (1) identical; meaning that taxon sampling has no effect on the time tree generation or (2) different; meaning that taxon sampling does have an effect on the timetree generation. If it is identical, then it should show a one-to-one relationship ($y = x$) with minimal residual variance (high R^2). We implemented the relative and absolute time estimation methodologies into the PATS pipeline to allow for eased replication and automation. Overall, we tested whether or not taxon sampling can lead to time tree discrepancies in those cases in which topology is not affected by facilitating the testing of molecular clock scenarios through our PATS pipeline.

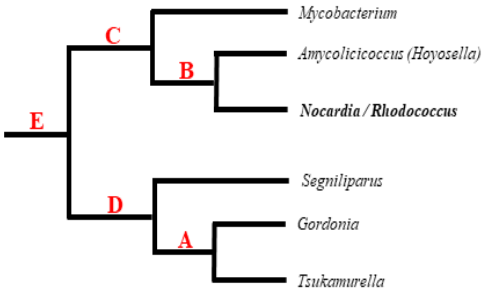
4.2.1 Implementation of RelTime into PATS……..RelTime has been implemented into the PATS pipeline to test the effects of taxon sampling on timetree estimates. Currently, MEGA version 11.0.13 has been implemented into the PATS pipeline and allows users to use it in two ways: (1) Timetree analyses via RelTime can be run without running the full PATS pipeline, or (2) time tree analyses can be added to a standard run of the PATS pipeline. Both of these require a concatenated sequence file (like the one generated by PATS concatenation step) and a tree topology file (a standard .nwk output tree from either FastTree or RAxML will work) as input files. Multiple settings within the configuration file (poPipeSettings.sh) will need to be set, such as model type, outgroup species, and others. Within the molecular clock subfolder of PATS, the user will be able to define the calibrations for that analysis in a control file (e.g., userPrimaryCalibrationInput.ctl). After the analysis is completed, a timetree will be

generated (divergence times) along with a summary file, relative time tree, and table of the divergence times.

4.2: Results and conclusions

4.2.1: Branch length……..Initially, we analyzed the branch length outputs for each of our four permutations (KeepOne) and the full tree from our 103 Actinobacteria analysis previously run through the PATS pipeline. The final trees were generated via RAxML with 100 bootstraps. We extracted each of the branch lengths that subtend each of the backbone nodes (A – E) (Fig. 8) for each of the permutations and full tree (Fig. 14). For each branch (A – E) of the Nocardia permutations, the branch lengths are within 1% of each other. When comparing the Nocardia to the Rhodococcus, we see a maximum of 2.5% difference in node A and a roughly 4.5% difference in node E. Since both the permutations with Nocardia and with Rhodococcus produce different topologies, the discrepancy in the deeper node E are expected: in the Nocardia permutations, the Mycobacterium lineage changes from being the sister to Nocardia/Amycolicicoccus to being the deepest branch in the topology and in the Rhodococcus permutations, Segniliparus changes from being sister to Gordonia/Tsukamurella to being the deepest branch in the topology. When comparing both the Nocardia and Rhodococcus to the full tree, we see similar trends. For branch lengths A and E, the Nocardia permutations are more similar to that of the full tree (maximum difference of 0.874 %). Branch B is relatively consistent across all species; however, it is slightly lower in the full tree. Overall, the branch lengths for all four of the permutations and the full tree do not differ enough to warrant a further analysis.

A.



B.

KeepOne Species	A_{branch}	B_{branch}	C_{branch}	D_{branch}	E_{branch}
Nocardia brasiliensis	0.06586	0.06205	*NA*	0.03500	0.06514
Nocardia cyriacigeorgic	0.06742	0.07405	*NA*	0.03836	0.07530
Nocardia nova	0.06249	0.06577	*NA*	0.03732	0.06688
Rhodococcus opacus	0.08692	0.07069	0.03628	*NA*	0.03068
Full Tree Branch Lengths	**0.06602**	**0.05116**	**0.03688**	**0.02957**	**0.06656**

Figure 14*:* Topologies and branch length supports of major groups in a phylogeny of 103 Actinobacteria. (**A.**) Topology of the genera Mycobacterium / Segniliparus / Gordonia / Tsukamurella / Nocardia / Rhodococcus from the full tree before the KeepOne permutations were applied. Each letter (A – E) represents a node that relates to (**B.**) to show the branch lengths and how the nodes change for each KeepOne scenario. In (**A.**) the bolded genera are those used during the permutations.

4.2.2: Divergence times……..Next, node times were estimated with RelTime, as implemented in MEGA, which produces relative and absolute time estimates [99]. For this analysis, a LG + G model was utilized with 5 gamma categories (as was done during the phylogenetic reconstruction process). Divergence time estimates were calculated via a maximum likelihood (heuristic) statistical method. This analysis was applied to two permutated trees, one with Nocardia (*Nocardia brasiliensis)* (Fig.15B) and one with Rhodococcus *(Rhodococcus opacus*) (Fig. 15C), and the full tree (Fig. 15A). These trees were selected because they have different topologies. Two calibration points were used, independently, one located close to the permutated clade (Mycobacterium - Nocardia/Rhodococcus) (Fig. 8; node C) and the other distant from the permutated clade (Segniliparus - Tsukamurella) (Fig. 8; node D). The divergence time estimates used as secondary calibrations were taken from the TimeTree database (TimeTree.org) [153]: *Mycobacterium* - *Nocardia/Rhodococcus* (M-N/R) [880 – 882 MYA] and *Segniliparus* - *Tsukamurella* (S-T) [958 - 960 MYA]. Time ranges were calculated by taking the divergence time estimates ± 1 MYA. We assumed a uniform distribution for each

calibrated node. By using a fixed secondary calibration estimate we were able to identify the source of time divergence variability to taxon choice or calibration choice instead of uncertainty in the calibration itself. Our divergence time estimations have focused only on secondary calibrations but, as shown previously by us [154], these can include large margins of error and biases. Thus, it will be necessary, in the future, to investigate the estimation of node ages using primary calibrations, if available. Moreover, when using secondary calibrations, we assumed a uniform distribution, but it is unknown if this assumption is unduly influencing the outcomes of our analyses. We rooted each of these trees with *Saccharothrix espanaensis, Kutzneria albida, Kibdelosporangium phytohabitans, Saccharomonospora viridis, Amycolatopsis methanolica, Amycolatopsis japonica, Amycolatopsis orientalis, Amycolatopsis mediterranei, Mycobacterium kansasii, Pseudonocardia sp. HH130629, Pseudonocardia dioxanivorans, Saccharopolyspora erythraea,* and *Actinosynnema mirum* as our outgroup species. A total of six timetrees were generated through RelTime and compared to each other with only shared nodes of our clade of interest (Fig. 14A) being utilized for direct comparisons.

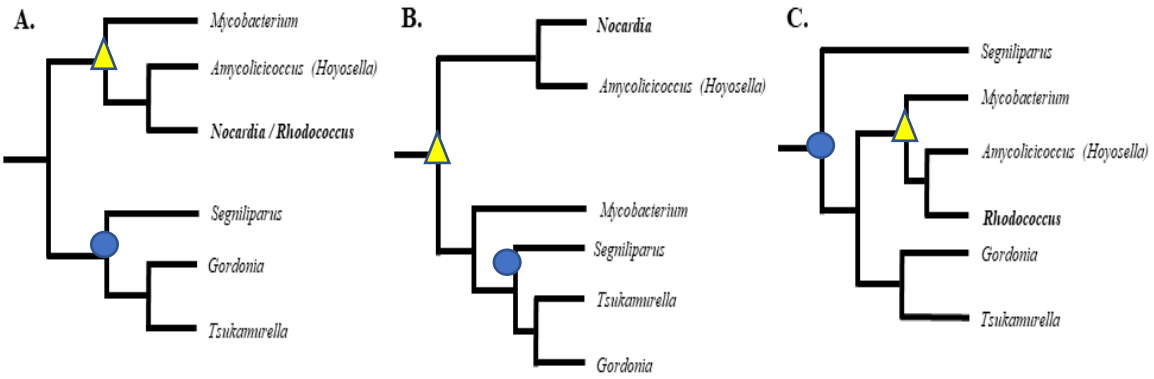


Figure 15: Tree topologies for each of the (**A.**) full tree, the (**B.**) Nocardia KeepOne tree, and the (**C.**) Rhodococcus KeepOne tree. Yellow triangle: location of the M-N/R calibration point. Blue circle: location of the S-T calibration point.

Initially, we compared RelTime divergence time estimates for the two tree topologies (Nocardia KeepOne scenario and Rhodococcus KeepOne scenario), using data generated from the two independent calibrations (M-N/R) and (S-T). When looking at the Nocardia topology (Nocardia (M-N/R) vs Nocardia (S-T)), we see that while the two calibrations produce time estimates that are linearly related to each other, the S-T calibration estimates node ages that are approximately 21% older (Fig. 16A). The Rhodococcus topology comparison of the two calibration shows the opposite effect, with the S-T calibration producing estimates slightly younger (6%) than the M-NR calibration (Fig. 16B). Lastly, for the full tree topology comparison the two calibrations produce estimates that are within 5% of each other (Fig. 16C). All comparisons have an R^2 of 1 which shows that the estimations from RelTime are consistent within each topology when similar datasets are used. The overestimation of the Nocardia (S-T) calibrated tree is likely caused by the depth and location of the Nocardia (M-N/R) node (Fig. 8; node C) due to its behavior when a Rhodococcus species is used. For the Rhodococcus trees, we see a similar outcome but in the opposite direction because the Rhodococcus (M-N/R) node is shallower in the tree which leads to older time estimates than does the Rhodococcus (T-S) calibrated node which flips this node to outside of the Mycobacterium clade (Fig. 8; node D). The full tree estimates are within 5% which shows that both calibrations performed relatively equally as both nodes exist in the full tree. These analyses show that calibration location within a topology is very important as it can change the timing of like-nodes even when the overall topology remains constant.

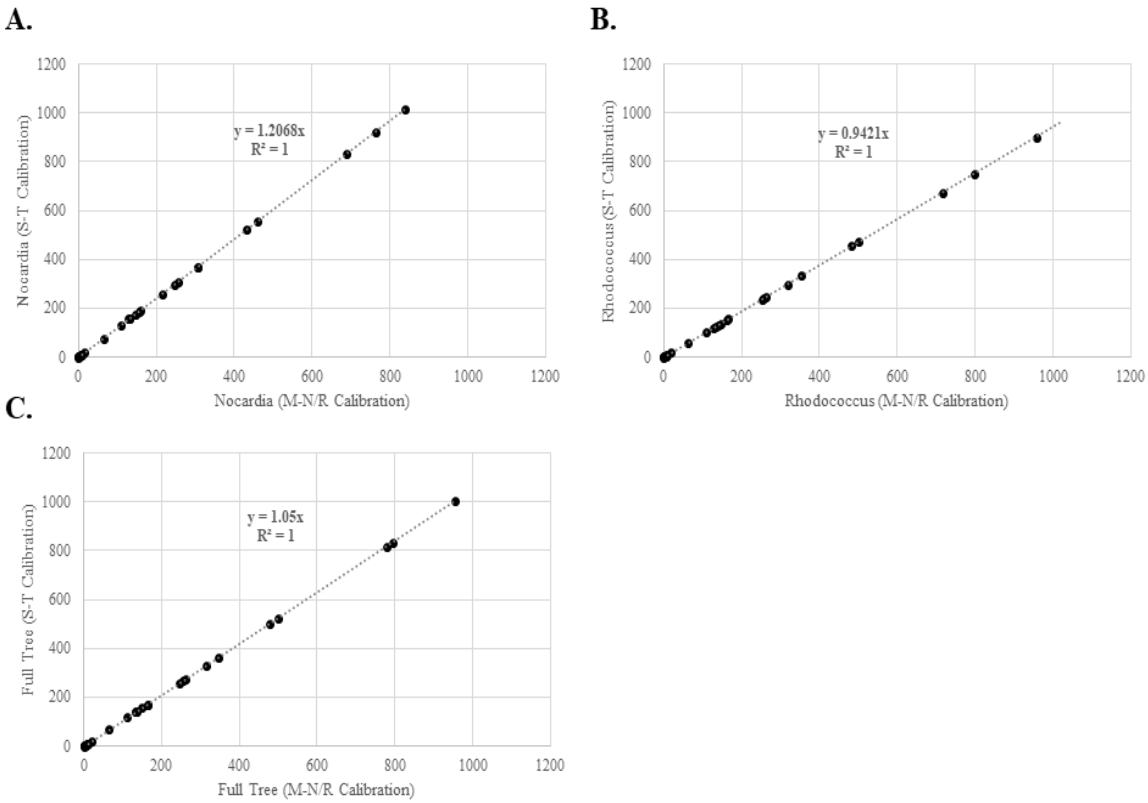


Figure 16: Divergence times plot estimated via RelTime for each Keepone scenario and full tree analysis. These plots show the relationship of the estimated divergence times for each of the calibrated nodes for the two Nocardia calibrated trees (**A.**), the two Rhodococcus calibrated trees (**B.**), and the full tree (**C.**). Each plot shows like-nodes from the Mycobacterium / Rhodococcus / Nocardia / Tsukamurella / Segniliparus / Gordonia clade. (S-T): Segniliparus-Tsukamurella; (M-N/R): Mycobacterium-Nocardia/Rhodococcus.

Next, we looked at the comparisons of the (S-T) calibrated nodes and the (M-N/R) calibrated nodes between the Nocardia KeepOne and the Rhodococcus KeepOne (Fig. 17A). We see that the Nocardia (S-T) calibrated tree has older time estimates compared to the Rhodococcus (S-T) calibrated tree. This is most likely caused by the change in the location of Segniliparus, which becomes the sister lineage of all species (Mycobacterium, Rhodococcus, Gordonia, Tsukamurella) rather than sister to only Gordonia/Tsukamurella as in the Nocardia-only permutation (Fig. 15). This node becoming deeper in the tree and being used as calibration affects the evolutionary rate of

every branch, thus leading to younger time estimates. For the M-N/R calibrated trees, we see the opposite result with the Rhodococcus time estimates being older than the Nocardia estimates because of the change in location of this calibration node (from shallower in the Rhodococcus permutation to deeper in the Nocardia permutation) (Figs. 15 and 17B). Overall, these analyses show the importance of species choice on the time tree estimates, especially when the affected nodes are used as calibration points.

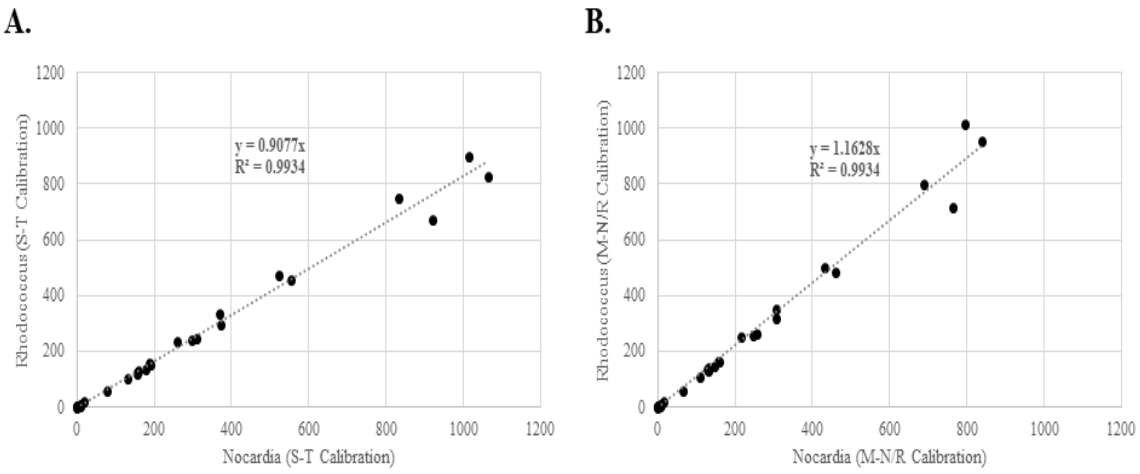


Figure 17: Divergence time plot estimated via RelTime for each calibrated node. These plots show the relationship of the estimated divergence times for each of the calibrated nodes. (**A.**) Segniliparus - Tsukamurella calibration and (**B.**) the Mycobacterium – Nocardia/Rhodococcus calibration.

Lastly, we looked at how each of the KeepOne permutations and full trees compared to each other with respect to divergence time estimates. In the cases in which the calibration was not placed on a node affected by taxon sampling we see little to no difference between estimations. Both the Rhodococcus (M-N/R) node (Fig. 18B) and the Nocardia (S-T) node (Fig. 18C) locations mimic that of the full tree topology, which leads to an overall difference of 1.54% and 6.87%, respectively. Instead, when calibrations are placed on nodes that change because of taxon sampling, we see results

similar to those discussed above, where the change in estimated times is either older (Fig. 18A) or younger (Fig. 18D) depending on the location of the calibrated node in the full tree. Even in these cases, however, it is still possible that divergence times could be included in the confidence intervals, which would decrease the severity of the discrepancies we are observing. We check this by comparing the CIs of the permutated trees to those of the full tree to check their respective sizes and inclusion of the estimated times. We find two trends: first, when the calibration is not affected by taxon sampling, the ranges of CIs are similar between permutation and full tree; second, for the vast majority of the nodes, the divergence times of the permutations are included within the CIs of the full tree, suggesting that, even when the times differ, they are within the variance of the estimates (Fig. A6).

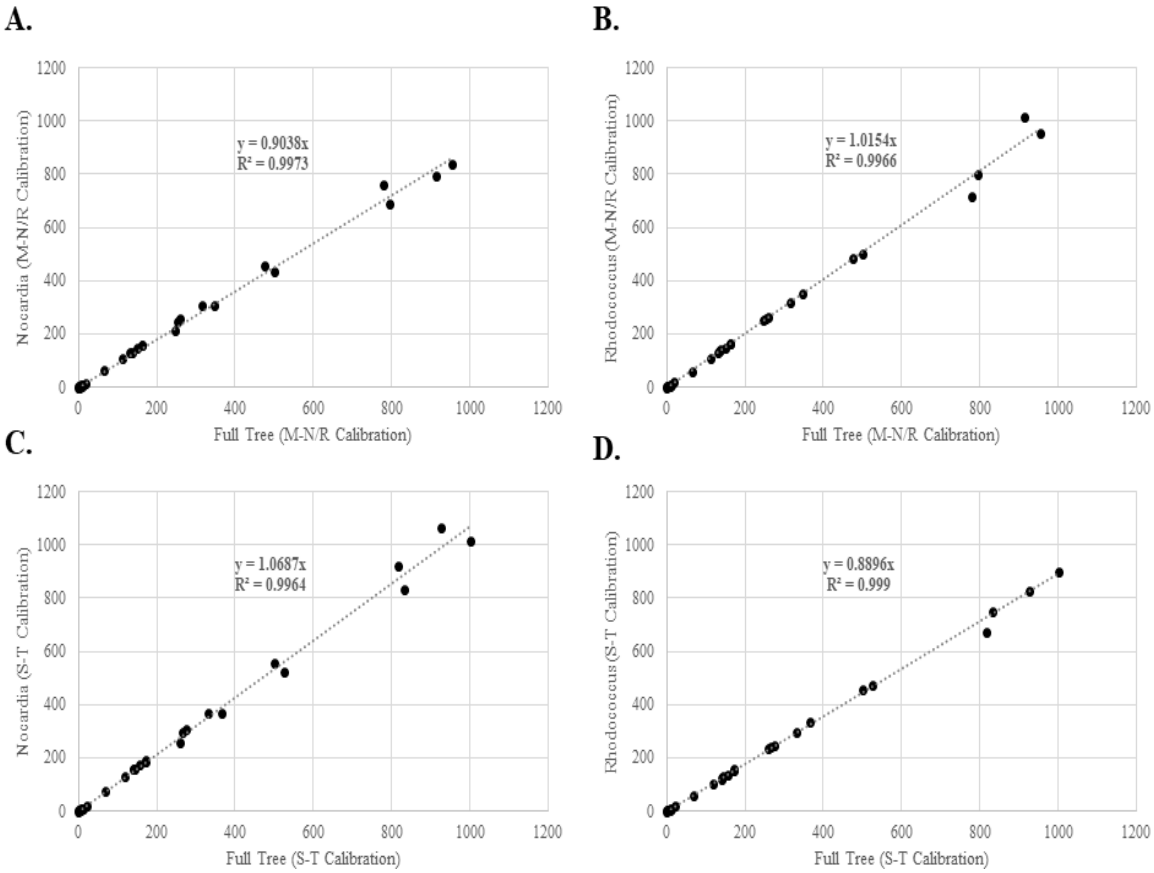


Figure 18: Divergence time plot estimated via RelTime for each permutation scenario versus the full trees. These plots show the relationship of the estimated divergence times for each of the permutations and full tree comparing like-calibrated nodes. (**A.**) Nocardia (M-N/R) calibrated tree versus the full tree (M-N/R) calibrated; (**B.**) Rhodococcus (M-N/R) calibrated tree versus the full tree (M-N/R) calibrated; (**C.**) Nocardia (S-T) calibrated tree versus the full tree (S-T) calibrated; (**D.**) Rhodococcus (S-T) calibrated tree versus the full tree (S-T) calibrated.

CHAPTER 5

CONCLUSIONS

In our initial branch length analysis, the branches seem to mostly be unaffected by taxon sampling, however, our timetree analyses show a different story. Because of the topological differences when either a Rhodococcus species or a Nocardia species is used, when compared to the full tree, the time estimates change, in particular when the nodes that are affected by the taxon sampling are the same nodes being calibrated. This shows the importance of testing the effects of taxon sampling before timetrees are generated. Thus, a tree reconstructed without considering the effects of taxon sampling could lead to a two-fold problem. First, the topology could change significantly, in our case based on either using the Rhodococcus or Nocardia species. Second, the estimated times for the origin of different groups could be substantially different, especially if the calibration node is directly affected by taxon sampling choices. However, point divergence times should not be considered as exact estimates but rather the confidence interval (CI) around them is a better estimate of the possible timeline for the origin of each lineage. When accounting for the uncertainty in time estimates provided by Cis, we find a very good agreement between time estimations in permutated and non-permutated trees. This is reassuring as it suggests that the discrepancies observed here caused by taxon sampling, even when large, might still allow to identify general evolutionary patterns. One downside of this approach, however, is that CI ranges can be large (hundreds of millions of years), thus leading to low precision of the reconstructed timelines.

Overall, the analyses in chapters 1, 2, and 3 show that taxon sampling is a key aspect of phylogenetic reconstructions, and it can affect topology and estimated times in various ways. The ever-increasing size of datasets used for evolutionary analyses is likely to lead to a more frequent use of taxon sampling especially because some steps of phylogenetic reconstructions still carry a heavy computational burden. Thus, evaluating the effects of the choices made when few species are used to represent large groups will become increasingly important to determine the accuracy of the results obtained. The new computational pipeline presented here, PATS, is a tool that enables the automatic testing of taxon sampling on phylogenetic and molecular time estimations without the need for tedious and error prone manual analyses. Moreover, these computational pipelines are streamlining bioinformatics analyses for those who wish to use the end results but are not familiar with the intricacies of procedures and protocols. This will have the effect of democratizing the use of bioinformatics by including research groups who may not have the expertise in bioinformatics but have the biological knowledge to carry out interpretations of the end results.

References

1. Darwin C (1859) On The Origin of Species by Means of Natural Selection, or Preservation of Favoured Races in the Struggle for Life. John Murray, London

2. Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6:361–375. https://doi.org/10.1038/nrg1603

3. Bailey SF, Blanquart F, Bataillon T, Kassen R (2017) What drives parallel evolution? BioEssays 39:e201600176. https://doi.org/10.1002/bies.201600176

4. Christin PA, Weinreich DM, Besnard G (2010) Causes and evolutionary significance of genetic convergence. Trends Genet 26:400–405. https://doi.org/10.1016/j.tig.2010.06.005

5. Jetz W, Thomas GH, Joy JB, et al (2012) The global diversity of birds in space and time. Nature 491:444–448. https://doi.org/10.1038/nature11631

6. Losos J (2013) The Princeton guide to evolution. Princeton University Press

7. R Vahdati A, Wagner A (2016) Parallel or convergent evolution in human population genomic data revealed by genotype networks. BMC Evol Biol 16:1–20. https://doi.org/10.1186/s12862-016-0722-0

8. Clerissi C, Touchon M, Capela D, et al (2018) Parallels between experimental and natural evolution of legume symbionts. Nat Commun 9:2264. https://doi.org/10.1038/s41467-018-04778-5

9. Zeller KA (1995) Phylogenetic relatedness within the genus Erysiphe estimated with morphological characteristics. Mycologia 87:525–531. https://doi.org/10.2307/3760771

10. Zamani Z, Shahi-Gharahlar A, Fatahi R, Bouzari N (2012) Genetic relatedness among some wild cherry (Prunus subgenus Cerasus) genotypes native to Iran assayed by morphological traits and random amplified polymorphic DNA analysis. Plant Syst Evol 298:499–509. https://doi.org/10.1007/s00606-011-0561-9

11. Sleator RD (2011) Phylogenetics. Arch Microbiol 193:235–239. https://doi.org/10.1007/s00203-011-0677-x

12. Lang JM, Darling AE, Eisen JA (2013) Phylogeny of Bacterial and Archaeal Genomes Using Conserved Genes: Supertrees and Supermatrices. PLoS One 8:e62510. https://doi.org/10.1371/journal.pone.0062510

13. Rinke C, Schwientek P, Sczyrba A, et al (2013) Insights into the phylogeny and coding potential of microbial dark matter. Nature 499:431–437. https://doi.org/10.1038/nature12352

14. Hug LA, Baker BJ, Anantharaman K, et al (2016) A new view of the tree of life. Nat Microbiol 1:16048. https://doi.org/10.1038/nmicrobiol.2016.48

15. Simion P, Philippe H, Baurain D, et al (2017) A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. Curr Biol 27:958–967. https://doi.org/10.1016/j.cub.2017.02.031

16. Linard B, Crampton-Platt A, Moriniere J, et al (2018) The contribution of mitochondrial metagenomics to large-scale data mining and phylogenetic analysis of Coleoptera. Mol Phylogenet Evol 128:1–11. https://doi.org/10.1016/j.ympev.2018.07.008

17. Shen H, Jin D, Shu JP, et al (2018) Large-scale phylogenomic analysis resolves a backbone phylogeny in ferns. Gigascience 7:1–11. https://doi.org/10.1093/gigascience/gix116

18. Estrada-Peña A, Cabezas-Cruz A (2019) Phyloproteomic and functional analyses do not support a split in the genus Borrelia (phylum Spirochaetes). BMC Evol Biol 19:54. https://doi.org/10.1186/s12862-019-1379-2

19. Shen H, Jin D, Shu J-P, et al (2018) Large-scale phylogenomic analysis resolves a backbone phylogeny in ferns. Gigascience 7:1–11. https://doi.org/10.1093/gigascience/gix116

20. Wang C, Gao Y, Lu B, et al (2021) Large-scale phylogenomic analysis provides new insights into the phylogeny of the class Oligohymenophorea (Protista, Ciliophora) with establishment of a new subclass Urocentria nov. subcl. Mol Phylogenet Evol 159:107112. https://doi.org/10.1016/j.ympev.2021.107112

21. Superson AA, Phelan D, Dekovich A, Battistuzzi FU (2019) Choice of species affects phylogenetic stability of deep nodes: An empirical example in Terrabacteria. Bioinformatics 35:3608–3616. https://doi.org/10.1093/bioinformatics/btz121

22. Ashkenazy H, Kliger Y (2010) Reducing phylogenetic bias in correlated mutation analysis. Protein Eng Des Sel 23:321–326. https://doi.org/10.1093/protein/gzp078

23. Duchêne DA, Duchêne S, Ho SYW (2017) New statistical criteria detect phylogenetic bias caused by compositional heterogeneity. Mol Biol Evol 34:1529–1534. https://doi.org/10.1093/molbev/msx092

24. Townsend JP, Su Z, Tekle YI (2012) Phylogenetic Signal and Noise: Predicting the Power of a Data Set to Resolve Phylogeny. Syst Biol 61:835. https://doi.org/10.1093/sysbio/sys036

25. Shen XX, Hittinger CT, Rokas A (2017) Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat Ecol Evol 1:0126. https://doi.org/10.1038/s41559-017-0126

26. Mongiardino Koch N, Gauthier JA (2018) Noise and biases in genomic data may underlie radically different hypotheses for the position of Iguania within Squamata. PLoS One 13:e0202729. https://doi.org/10.1371/journal.pone.0202729

27. Rokas A, Carroll SB (2005) More Genes or More Taxa? The Relative Contribution of Gene Number and Taxon Number to Phylogenetic Accuracy. Mol Biol Evol 22:1337–1344. https://doi.org/10.1093/molbev/msi121

28. Sperling EA, Peterson KJ, Pisani D (2009) Phylogenetic-Signal Dissection of Nuclear Housekeeping Genes Supports the Paraphyly of Sponges and the Monophyly of Eumetazoa. Mol Biol Evol 26:2261–2274. https://doi.org/10.1093/molbev/msp148

29. Kumar S, Filipski AJ, Battistuzzi FU, et al (2012) Statistics and Truth in Phylogenomics. Mol Biol Evol 29:457–472. https://doi.org/10.1093/molbev/msr202

30. Pisani D, Pett W, Dohrmann M, et al (2015) Genomic data do not support comb jellies as the sister group to all other animals. Proc Natl Acad Sci U S A 112:15402–15407. https://doi.org/10.1073/pnas.1518127112

31. Hejase HA, Liu KJ (2016) A scalability study of phylogenetic network inference methods using empirical datasets and simulations involving a single reticulation. BMC Bioinformatics 17:422. https://doi.org/10.1186/s12859-016-1277-1

32. Fourment M, Magee AF, Whidden C, et al (2020) 19 Dubious Ways to Compute the Marginal Likelihood of a Phylogenetic Tree Topology. Syst Biol 69:209–220. https://doi.org/10.1093/sysbio/syz046

33. Whidden C, Claywell BC, Fisher T, et al (2020) Systematic Exploration of the High Likelihood Set of Phylogenetic Tree Topologies. Syst Biol 69:280–293. https://doi.org/10.1093/sysbio/syz047

34. Rodríguez A, Burgon JD, Lyra M, et al (2017) Inferring the shallow phylogeny of true salamanders (Salamandra) by multiple phylogenomic approaches. Mol Phylogenet Evol 115:16–26. https://doi.org/10.1016/j.ympev.2017.07.009

35. Joyce J (2019) Bayes' Theorem. In: Zalta EN (ed) The {Stanford} Encyclopedia of Philosophy, Spring 201. Metaphysics Research Lab, Stanford University

36. Venn J (1888) The Logic of Chance. 3rd ed, Macmillan co, Repr New York

37. Neyman J (1937) Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. Philos Trans R Soc London Ser A, Math Phys Sci 236:333–380. https://doi.org/10.1098/rsta.1937.0005

38. McCormack GP, Clewley JP (2002) The application of molecular phylogenetics to the analysis of viral genome diversity and evolution. Rev Med Virol 12:221–238. https://doi.org/10.1002/rmv.355

39. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454

40. Sokal RR (1958) A statistical method for evaluating systematic relationships

41. Ané C, Larget B, Baum DA, et al (2007) Bayesian estimation of concordance among gene trees. Mol Biol Evol 24:412–426. https://doi.org/10.1093/molbev/msl170

42. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7:214. https://doi.org/10.1186/1471-2148-7-214

43. Ronquist F, Teslenko M, van der Mark P, et al (2012) MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Syst Biol 61:539–542. https://doi.org/10.1093/sysbio/sys029

44. Bouckaert R, Vaughan TG, Barido-Sottani J, et al (2019) BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLOS Comput Biol 15:e1006650. https://doi.org/10.1371/journal.pcbi.1006650

45. Price MN, Dehal PS, Arkin AP (2009) FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. Mol Biol Evol 26:1641–1650. https://doi.org/10.1093/molbev/msp077

46. Price MN, Dehal PS, Arkin AP (2010) FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS One 5:e9490. https://doi.org/10.1371/journal.pone.0009490

47. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033

48. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol 32:268–274. https://doi.org/10.1093/molbev/msu300

49. Minh BQ, Schmidt HA, Chernomor O, et al (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol 37:1530–1534. https://doi.org/10.1093/molbev/msaa015

50. Hoang DT, Vinh LS, Flouri T, et al (2018) MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. BMC Evol Biol 18:11. https://doi.org/10.1186/s12862-018-1131-3

51. Giribet G (2007) Efficient tree searches with available algorithms. Evol Bioinforma 3:341–356. https://doi.org/10.1177/117693430700300014

52. Sul SJ, Matthews S, Williams TL (2009) Using tree diversity to compare phylogenetic heuristics. BMC Bioinformatics 10:1–9. https://doi.org/10.1186/1471-2105-10-S4-S3

53. Davison A, D H (1997) Bootstrap methods and their application. Cambridge University Press

54. Stamatakis A, Hoover P, Rougemont J (2008) A Rapid Bootstrap Algorithm for the RAxML Web Servers. Syst Biol 57:758–771. https://doi.org/10.1080/10635150802429642

55. Nei M, Kumar S (2000) Molecular evolution and phylogenetics. Oxford University Press

56. Kapli P, Yang Z, Telford MJ (2020) Phylogenetic tree building in the genomic age. Nat Rev Genet 21:428–444. https://doi.org/10.1038/s41576-020-0233-0

57. JUKES TH, CANTOR CR (1969) Evolution of Protein Molecules. In: Mammalian Protein Metabolism. Elsevier, pp 21–132

58. Dayhoff MO (1969) Atlas of protein sequence and structure. National Biomedical Research Foundation.

59. Lanave C, Preparata G, Sacone C, Serio G (1984) A new method for calculating evolutionary substitution rates. J Mol Evol 20:86–93. https://doi.org/10.1007/BF02101990

60. Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. Mol Biol Evol 25:1307–1320. https://doi.org/10.1093/molbev/msn067

61. Gatto L, Catanzaro D, Milinkovitch MC (2006) Assessing the Applicability of the GTR Nucleotide Substitution Model through Simulations. Evol Bioinforma 2:117693430600200. https://doi.org/10.1177/117693430600200020

62. Le SQ, Dang CC, Gascuel O (2012) Modeling protein evolution with several amino acid replacement matrices depending on site rates. Mol Biol Evol 29:2921–2936. https://doi.org/10.1093/molbev/mss112

63. Lopez P, Casane D, Philippe H (2002) Heterotachy, an important process of protein evolution. Mol Biol Evol 19:1–7. https://doi.org/10.1093/oxfordjournals.molbev.a003973

64. Lartillot N, Philippe H (2004) A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. Mol Biol Evol 21:1095–1109. https://doi.org/10.1093/molbev/msh112

65. Pick KS, Philippe H, Schreiber F, et al (2010) Improved Phylogenomic Taxon Sampling Noticeably Affects Nonbilaterian Relationships. Mol Biol Evol 27:1983–1987. https://doi.org/10.1093/molbev/msq089

66. Plazzi F, Ferrucci RR, Passamonti M (2010) Phylogenetic representativeness: a new method for evaluating taxon sampling in evolutionary studies. BMC Bioinformatics 11:209. https://doi.org/10.1186/1471-2105-11-209

67. Esselstyn JA, Oliveros CH, Swanson MT, Faircloth BC (2017) Investigating Difficult Nodes in the Placental Mammal Tree with Expanded Taxon Sampling and Thousands of Ultraconserved Elements. Genome Biol Evol 9:2308–2321. https://doi.org/10.1093/gbe/evx168

68. Park DS, Worthington S, Xi Z (2018) Taxon sampling effects on the quantification and comparison of community phylogenetic diversity. Mol Ecol 27:1296–1308. https://doi.org/10.1111/mec.14520

69. Zou H, Jakovlić I, Zhang D, et al (2020) Architectural instability, inverted skews and mitochondrial phylogenomics of Isopoda: outgroup choice affects the long-branch attraction artefacts. R Soc Open Sci 7:191887. https://doi.org/10.1098/rsos.191887

70. Mukherjee S, Stamatis D, Bertsch J, et al (2019) Genomes OnLine database (GOLD) v.7: Updates and new features. Nucleic Acids Res 47:D649–D659. https://doi.org/10.1093/nar/gky977

71. Mukherjee S, Stamatis D, Bertsch J, et al (2021) Genomes OnLine Database (GOLD) v.8: overview and updates. Nucleic Acids Res 49:D723–D733. https://doi.org/10.1093/nar/gkaa983

72. Wheeler WC, Coddington JA, Crowley LM, et al (2017) The spider tree of life: phylogeny of Araneae based on target-gene analyses from an extensive taxon sampling. Cladistics 33:574–616. https://doi.org/10.1111/cla.12182

73. Çıplak B, Yahyaoğlu Ö, Uluar O (2021) Revisiting Pholidopterini (Orthoptera, Tettigoniidae): Rapid radiation causes homoplasy and phylogenetic instability. Zool Scr 50:225–240. https://doi.org/10.1111/zsc.12463

74. Żyła D, Bogri A, Heath TA, Solodovnikov A (2021) Total-evidence analysis resolves the phylogenetic position of an enigmatic group of Paederinae rove beetles (Coleoptera: Staphylinidae). Mol Phylogenet Evol 157:107059. https://doi.org/10.1016/j.ympev.2020.107059

75. Linder CR, Suri R, Liu K, Warnow T (2010) Benchmark datasets and software for developing and testing methods for large-scale multiple sequence alignment and phylogenetic inference. PLoS Curr 2:RRN1195. https://doi.org/10.1371/currents.RRN1195

76. Didelot X, Croucher NJ, Bentley SD, et al (2018) Bayesian inference of ancestral dates on bacterial phylogenetic trees. Nucleic Acids Res 46:e134–e134. https://doi.org/10.1093/nar/gky783

77. Lees JA, Kendall M, Parkhill J, et al (2018) Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. Wellcome Open Res 3:33. https://doi.org/10.12688/wellcomeopenres.14265.2

78. Wang Q (2019) Benchmarking and comparing software for phylogenetic analysis from genome-scale data

79. Li L (2003) OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res 13:2178–2189. https://doi.org/10.1101/gr.1224503

80. Chen F (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res 34:D363–D368. https://doi.org/10.1093/nar/gkj123

81. Zerbino DR, Achuthan P, Akanni W, et al (2018) Ensembl 2018. Nucleic Acids Res 46:D754–D761. https://doi.org/10.1093/nar/gkx1098

82. Lechner M, Findeiß S, Steiner L, et al (2011) Proteinortho: Detection of (Co-)orthologs in large-scale analysis. BMC Bioinformatics 12:124. https://doi.org/10.1186/1471-2105-12-124

83. Sonnhammer ELL, Östlund G (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic Acids Res 43:D234–D239. https://doi.org/10.1093/nar/gku1203

84. Huerta-Cepas J, Szklarczyk D, Heller D, et al (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 47:D309–D314. https://doi.org/10.1093/nar/gky1085

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

86. Notredame C, Higgins DG, Heringa J (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment 1 1Edited by J. Thornton. J Mol Biol 302:205–217. https://doi.org/10.1006/jmbi.2000.4042

87. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340

88. Sievers F, Higgins DG (2018) Clustal Omega for making accurate alignments of many protein sequences. Protein Sci 27:135–145. https://doi.org/10.1002/pro.3290

89. Katoh K, Rozewicki J, Yamada KD (2019) MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform 20:1160–1166. https://doi.org/10.1093/bib/bbx108

90. de Queiroz A, Gatesy J (2007) The supermatrix approach to systematics. Trends Ecol Evol 22:34–41. https://doi.org/10.1016/j.tree.2006.10.002

91. Cotton JA, Wilkinson M (2009) Supertrees join the mainstream of phylogenetics. Trends Ecol Evol 24:1–3. https://doi.org/10.1016/j.tree.2008.08.006

92. Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol 24:1586–1591. https://doi.org/10.1093/molbev/msm088

93. Larget BR, Kotha SK, Dewey CN, Ané C (2010) BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics 26:2910–2911. https://doi.org/10.1093/bioinformatics/btq539

94. Wilson IJ, Weale ME, Balding DJ (2003) Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. J R Stat Soc Ser A (Statistics Soc 166:155–188. https://doi.org/10.1111/1467-985X.00264

95. Thomas GH, Hartmann K, Jetz W, et al (2013) PASTIS: an R package to facilitate phylogenetic assembly with soft taxonomic inferences. Methods Ecol Evol 4:1011–1017. https://doi.org/10.1111/2041-210X.12117

96. Goloboff PA, Farris JS, Nixon KC (2008) TNT, a free program for phylogenetic analysis. Cladistics 24:774–786. https://doi.org/10.1111/j.1096-0031.2008.00217.x

97. Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17:1244–1245. https://doi.org/10.1093/bioinformatics/17.12.1244

98. Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol 33:1870–1874. https://doi.org/10.1093/molbev/msw054

99. Kumar S, Stecher G, Li M, et al (2018) MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 35:1547–1549. https://doi.org/10.1093/molbev/msy096

100. Swofford D (2002) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0b10

101. Howe K, Bateman A, Durbin R (2002) QuickTree: building huge Neighbour-Joining trees of protein sequences. Bioinformatics 18:1546–1547. https://doi.org/10.1093/bioinformatics/18.11.1546

102. Vinh LS, Von Haeseler A (2004) IQPNNI: Moving fast through tree space and stopping in time. Mol Biol Evol 21:1565–1571. https://doi.org/10.1093/molbev/msh176

103. Letunic I, Bork P (2019) Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 47:W256–W259. https://doi.org/10.1093/nar/gkz239

104. James TY, Stajich JE, Hittinger CT, Rokas A (2020) Toward a Fully Resolved Fungal Tree of Life. Annu Rev Microbiol 74:291–313. https://doi.org/10.1146/annurev-micro-022020-051835

105. Williams TA, Cox CJ, Foster PG, et al (2020) Phylogenomics provides robust support for a two-domains tree of life. Nat Ecol Evol 4:138–147. https://doi.org/10.1038/s41559-019-1040-x

106. Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60. https://doi.org/10.1038/nmeth.3176

107. Buchfink B, Reuter K, Drost H (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods 18:. https://doi.org/10.1038/s41592-021-01101-x

108. Camacho C, Coulouris G, Avagyan V, et al (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421

109. Altschul SF, Gish W, Miller W, et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

110. Altschul SF, Madden TL, Schäffer AA, et al (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. https://doi.org/10.1093/nar/25.17.3389

111. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci 89:10915–10919. https://doi.org/10.1073/pnas.89.22.10915

112. Afgan E, Nekrutenko A, Grüning BA, et al (2022) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. Nucleic Acids Res 50:W345–W351. https://doi.org/10.1093/nar/gkac247

113. Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. Math Biosci 53:131–147. https://doi.org/10.1016/0025-5564(81)90043-2

114. Shimodaira H, Hasegawa M (1999) Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. Mol Biol Evol 16:1114–1116. https://doi.org/10.1093/oxfordjournals.molbev.a026201

115. Shimodaira H (2002) An Approximately Unbiased Test of Phylogenetic Tree Selection. Syst Biol 51:492–508. https://doi.org/10.1080/10635150290069913

116. Nouioui I, Carro L, García-López M, et al (2018) Genome-Based Taxonomic Classification of the Phylum Actinobacteria. Front Microbiol 9:1–119. https://doi.org/10.3389/fmicb.2018.02007

117. Salam N, Jiao J-Y, Zhang X-T, Li W-J (2020) Update on the classification of higher ranks in the phylum Actinobacteria. Int J Syst Evol Microbiol 70:1331–1355. https://doi.org/10.1099/ijsem.0.003920

118. Yamamoto T, Hasegawa Y, Lau PCK, Iwaki H (2021) Identification and characterization of a chc gene cluster responsible for the aromatization pathway of cyclohexanecarboxylate degradation in Sinomonas cyclohexanicum ATCC 51369. J Biosci Bioeng 132:621–629. https://doi.org/10.1016/j.jbiosc.2021.08.013

119. Marjanović D, Laurin M (2007) Fossils, Molecules, Divergence Times, and the Origin of Lissamphibians. Syst Biol 56:369–388. https://doi.org/10.1080/10635150701397635

120. Marjanović D, Laurin M (2008) Assessing Confidence Intervals for Stratigraphic Ranges of Higher Taxa: The Case of Lissamphibia. Acta Palaeontol Pol 53:413–432. https://doi.org/10.4202/app.2008.0305

121. Ho SYW, Phillips MJ (2009) Accounting for Calibration Uncertainty in Phylogenetic Estimation of Evolutionary Divergence Times. Syst Biol 58:367–380. https://doi.org/10.1093/sysbio/syp035

122. Inoue J, Donoghue PCJ, Yang Z (2010) The Impact of the Representation of Fossil Calibrations on Bayesian Estimation of Species Divergence Times. Syst Biol 59:74–89. https://doi.org/10.1093/sysbio/syp078

123. Sauquet H, Ho SYW, Gandolfo MA, et al (2012) Testing the Impact of Calibration on Molecular Divergence Times Using a Fossil-Rich Group: The Case of Nothofagus (Fagales). Syst Biol 61:289–313. https://doi.org/10.1093/sysbio/syr116

124. Sterli J, Pol D, Laurin M (2013) Incorporating phylogenetic uncertainty on phylogeny-based palaeontological dating and the timing of turtle diversification. Cladistics 29:233–246. https://doi.org/10.1111/j.1096-0031.2012.00425.x

125. Heath TA, Huelsenbeck JP, Stadler T (2014) The fossilized birth-death process for coherent calibration of divergence-time estimates. Proc Natl Acad Sci 111:E2957–E2966. https://doi.org/10.1073/pnas.1319091111

126. Hipsley CA, Müller J (2014) Beyond fossil calibrations: realities of molecular clock practices in evolutionary biology. Front Genet 5:138. https://doi.org/10.3389/fgene.2014.00138

127. Warnock RCM, Parham JF, Joyce WG, et al (2015) Calibration uncertainty in molecular dating analyses: there is no substitute for the prior evaluation of time priors. Proc R Soc B Biol Sci 282:20141013. https://doi.org/10.1098/rspb.2014.1013

128. dos Reis M, Donoghue PCJ, Yang Z (2016) Bayesian molecular clock dating of species divergences in the genomics era. Nat Rev Genet 17:71–80. https://doi.org/10.1038/nrg.2015.8

129. Kumar S, Hedges SB (2016) Advances in Time Estimation Methods for Molecular Data. Mol Biol Evol 33:863–869. https://doi.org/10.1093/molbev/msw026

130. Didier G, Fau M, Laurin M (2017) Likelihood of Tree Topologies with Fossils and Diversification Rate Estimation. Syst Biol 66:964–987. https://doi.org/10.1093/sysbio/syx045

131. Bromham L (2019) Six Impossible Things before Breakfast: Assumptions, Models, and Belief in Molecular Dating. Trends Ecol Evol 34:474–486. https://doi.org/10.1016/j.tree.2019.01.017

132. Marshall CR (2019) Using the Fossil Record to Evaluate Timetree Timescales. Front Genet 10:1–20. https://doi.org/10.3389/fgene.2019.01049

133. Ho SYM (2007) Calibrating molecular estimates of substitution rates and divergence times in birds. J Avian Biol 38:409–414. https://doi.org/10.1111/j.0908-8857.2007.04168.x

134. Magnabosco C, Moore KR, Wolfe JM, Fournier GP (2018) Dating phototrophic microbial lineages with reticulate gene histories. Geobiology 16:179–189. https://doi.org/10.1111/gbi.12273

135. Wolfe JM, Fournier GP (2018) Horizontal gene transfer constrains the timing of methanogen evolution. Nat Ecol Evol 2:897–903. https://doi.org/10.1038/s41559-018-0513-7

136. Britton T (2005) Estimating Divergence Times in Phylogenetic Trees Without a Molecular Clock. Syst Biol 54:500–507. https://doi.org/10.1080/10635150590947311

137. Hug LA, Roger AJ (2007) The Impact of Fossils and Taxon Sampling on Ancient Molecular Dating Analyses. Mol Biol Evol 24:1889–1897. https://doi.org/10.1093/molbev/msm115

138. Perie MD, Doyle JA (2012) Dating clades with fossils and molecules: the case of Annonaceae. Bot J Linn Soc 169:84–116. https://doi.org/10.1111/j.1095-8339.2012.01234.x

139. dos Reis M, Yang Z (2013) The unbearable uncertainty of Bayesian divergence time estimation. J Syst Evol 51:30–43. https://doi.org/10.1111/j.1759-6831.2012.00236.x

140. Marshall CR (2008) A Simple Method for Bracketing Absolute Divergence Times on Molecular Phylogenies Using Multiple Fossil Calibration Points. Am Nat 171:726–742. https://doi.org/10.1086/587523

141. Warnock RCM, Yang Z, Donoghue PCJ (2017) Testing the molecular clock using mechanistic models of fossil preservation and molecular evolution. Proc R Soc B Biol Sci 284:20170227. https://doi.org/10.1098/rspb.2017.0227

142. Schenk JJ (2016) Consequences of Secondary Calibrations on Divergence Time Estimates. PLoS One 11:e0148228. https://doi.org/10.1371/journal.pone.0148228

143. Graur D, Martin W (2004) Reading the entrails of chickens: Molecular timescales of evolution and the illusion of precision. Trends Genet 20:80–86. https://doi.org/10.1016/j.tig.2003.12.003

144. Reisz RR, Müller J (2004) Molecular timescales and the fossil record: a paleontological perspective. Trends Genet 20:237–241. https://doi.org/10.1016/j.tig.2004.03.007

145. Tamura K, Battistuzzi FU, Billing-Ross P, et al (2012) Estimating divergence times in large molecular phylogenies. Proc Natl Acad Sci U S A 109:19333–19338. https://doi.org/10.1073/pnas.1213199109

146. Tamura K, Tao Q, Kumar S (2018) Theoretical Foundation of the RelTime Method for Estimating Divergence Times from Variable Evolutionary Rates. Mol Biol Evol 35:1770–1782. https://doi.org/10.1093/molbev/msy044

147. Hedges SB, Kumar S (2009) The Timetree of Life, 1st ed. Oxford University Press

148. Thorne JL, Kishino H (2002) Divergence Time and Evolutionary Rate Estimation with Multilocus Data. Syst Biol 51:689–702. https://doi.org/10.1080/10635150290102456

149. Rosenberg MS, Kumar S (2003) Heterogeneity of Nucleotide Frequencies Among Evolutionary Lineages and Phylogenetic Inference. Mol Biol Evol 20:610–621. https://doi.org/10.1093/molbev/msg067

150. Grassly NC, Adachj J, Rambaut A (1997) PSeq-Gen: an application for the Monte Carlo simulation of protein sequence evolution along phylogenetic trees. Bioinformatics 13:559–560. https://doi.org/10.1093/bioinformatics/13.5.559

151. Brochu CA (2004) Patterns of Calibration Age Sensitivity With Quartet Dating Methods. J Paleontol 78:7–30. https://doi.org/10.1666/0022-3360(2004)078<0007:pocasw>2.0.co;2

152. Tamura K, Stecher G, Kumar S (2021) MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Mol Biol Evol 38:3022–3027. https://doi.org/10.1093/molbev/msab120

153. Kumar S, Stecher G, Suleski M, Hedges SB (2017) TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol 34:1812–1819. https://doi.org/10.1093/molbev/msx116

154. Powell CLE, Waskin S, Battistuzzi FU (2020) Quantifying the Error of Secondary vs. Distant Primary Calibrations in a Simulated Environment. Front Genet 11:1–9. https://doi.org/10.3389/fgene.2020.00252

APPENDIX

Table A1: Actinobacteria species that were a part of the 103 dataset. *N/A depicts entry that is no longer available in NCBI database.*

103 Actinobacteria Dataset			
Genus	**Species**	**Strain**	**GCF / GCA #**
Actinosynnema	mirum	DSM_43827	GCF_000023245.1
Amycolatopsis	japonica	DSM 44213	GCF_000732925.1
Amycolatopsis	lurida	NRRL_2430	GCF_000749465.2
Amycolatopsis	mediterranei	S699	GCF_000220945.1
Amycolatopsis	methanolica	239	GCF_000371885.1
Amycolatopsis	orientalis	HCCB10007	GCA_000400635.2
Amycolicicoccus	subflavus	DQS3_9A1	GCF_000214175.1
Corynebacteriales	bacterium	X1036	GCF_001293125.1
Corynebacteriales	bacterium	X1698	GCF_001281505.1
Corynebacterium	argentoratense	DSM_44202	GCF_000590555.1
Corynebacterium	atypicum	R2070	GCF_000732945.1
Corynebacterium	aurimucosum	ATCC_700975	GCF_000022905.1
Corynebacterium	callunae	DSM_20147	GCF_000344785.1
Corynebacterium	camporealensis	DSM 44610	GCF_000980815.1
Corynebacterium	casei	LMG_S_19264	GCF_000550785.1
Corynebacterium	deserti	GIMN1.10	GCF_001277995.1
Corynebacterium	diphtheriae	31A	GCF_000241875.1
Corynebacterium	doosanense	CAU_212 = DSM_45436	GCF_000767055.1
Corynebacterium	efficiens	YS_314	GCF_000011305.1
Corynebacterium	epidermidicanis	DSM 45586	GCF_001021025.1
Corynebacterium	falsenii	DSM_44353	GCF_000525655.1
Corynebacterium	glutamicum	R	GCF_000010225.1
Corynebacterium	glyciniphilum	AJ_3170	GCF_000626675.1
Corynebacterium	halotolerans	YIM_70093 = DSM_44683	GCF_000341345.1
Corynebacterium	humireducens	NBRC_106098 = DSM_45392	GCF_000819445.1
Corynebacterium	imitans	DSM 44264	GCA_000739455.1
Corynebacterium	jeikeium	K411	GCF_000006605.1
Corynebacterium	kroppenstedtii	DSM_44385	GCF_000023145.1
Corynebacterium	kutscheri	DSM 20755	GCF_000980835.1
Corynebacterium	lactis	RW2_5	GCF_001274895.1
Corynebacterium	marinum	DSM_44953	GCF_000835165.1
Corynebacterium	maris	DSM_45190	GCF_000442645.1

Corynebacterium	mustelae	DSM 45274	GCF_001020985.1
Corynebacterium	pseudotuberculosis	1/06-A	GCA_000233735.1
Corynebacterium	resistens	DSM_45100	GCF_000177535.2
Corynebacterium	singulare	IBS B52218	GCF_000833575.1
Corynebacterium	sp.	ATCC_6931	GCF_000755185.1
Corynebacterium	stationis	622=DSM 20302	GCA_001941345.1
Corynebacterium	terpenotabidum	Y_11	GCF_000418365.1
Corynebacterium	testudinoris	DSM 44614	GCF_001021045.1
Corynebacterium	ulcerans	FRC58	GCF_000499805.2
Corynebacterium	urealyticum	DSM_7109	GCF_000069945.1
Corynebacterium	ureicelerivorans	IMMIB RIV-2301	GCF_000747315.1
Corynebacterium	uterequi	DSM 45634	GCF_001021065.1
Corynebacterium	variabile	DSM_44702	GCF_000179395.2
Corynebacterium	vitaeruminis	DSM_20294	GCF_000550805.1
Gordonia	bronchialis	DSM_43247	GCF_000024785.1
Gordonia	polyisoprenivorans	VH2	GCF_000247715.1
Gordonia	sp.	KTR9	GCF_000143885.2
Gordonia	sp.	QH_11	GCF_001305675.1
Kibdelosporangium	phytohabitans	KLBMP1111	GCF_001302585.1
Kutzneria	albida	DSM_43870	GCF_000525635.1
Mycobacterium	abscessus	subsp._bolletii_50594	GCF_000445035.1
Mycobacterium	africanum	GM041182	GCA_000253355.1
Mycobacterium	avium	104	GCF_000014985.1
Mycobacterium	bovis	BCG_3281	GCA_001043255.1
Mycobacterium	chubuense	NBB4	GCF_000266905.1
Mycobacterium	fortuitum	CT6	GCA_001307545.1
Mycobacterium	gilvum	PYR_GCK	GCA_025821885.1
Mycobacterium	goodii	ATCC 700504	GCF_022370755.2
Mycobacterium	haemophilum	DSM_44634	GCF_000340435.2
Mycobacterium	indicus_pranii	MTCC_9506	GCF_000298095.1
Mycobacterium	intracellulare	MOTT_64	GCF_000276825.1
Mycobacterium	kansasii	ATCC_12478	GCF_000157895.3
Mycobacterium	leprae	Br4923	GCF_000026685.1
Mycobacterium	liflandii	128FXT	GCF_000026445.2
Mycobacterium	marinum	M	GCF_000018345.1
Mycobacterium	microti	12	GCA_001544815.1
Mycobacterium	neoaurum	VKM_Ac_1815D	GCF_000317305.3
Mycobacterium	rhodesiae	NBB3	GCF_000230895.2
Mycobacterium	sinense	JDM601	GCF_000214155.1
Mycobacterium	sp.	EPa45	GCF_001021385.1
Mycobacterium	sp.	JLS	GCA_000016005.1
Mycobacterium	sp.	JS623	GCF_000328565.1

Mycobacterium	sp.	KMS	N/A
Mycobacterium	sp.	MCS	GCA_000014165.1
Mycobacterium	sp.	MOTT36Y	GCF_000262165.1
Mycobacterium	sp.	VKM_Ac_1817D	GCA_000416365.2
Mycobacterium	tuberculosis	H37Rv	GCA_000195955.2
Mycobacterium	ulcerans	Agy99	GCF_000013925.1
Mycobacterium	vanbaalenii	PYR_1	GCF_000015305.1
Mycobacterium	yongonense	05_1390	GCF_000418535.1
Nocardia	brasiliensis	ATCC_700358	GCF_000250675.2
Nocardia	cyriacigeorgica	GUH_2	GCF_000284035.1
Nocardia	farcinica	IFM 10152	GCA_000009805.1
Nocardia	nova	SH22a	GCF_000523235.1
Pseudonocardia_	dioxanivorans	CB1190	GCF_000196675.2
Pseudonocardia	sp.	AL041005_10	GCF_001294605.1
Pseudonocardia	sp.	EC080610_09	GCF_001420975.1
Pseudonocardia	sp.	EC080619_01	GCF_001420995.1
Pseudonocardia	sp.	EC080625_04	GCF_001294425.1
Pseudonocardia	sp.	HH130629_09	GCF_001294645.1
Rhodococcus	erythropolis	PR4	GCF_000010105.1
Rhodococcus	jostii	RHA1	GCF_000014565.1
Rhodococcus	opacus	PD630	GCF_020542785.1
Rhodococcus	pyridinivorans	SB3094	GCF_000511305.1
Rhodococcus	sp.	B7740	GCF_000954115.1
Saccharomonospora	viridis	DSM_43017	GCF_000023865.1
Saccharopolyspora	erythraea	NRRL_2338	GCF_000062885.1
Saccharothrix	espanaensis	DSM_44229	GCF_000328705.1
Segniliparus	rotundus	DSM_44985	GCF_000092825.1
Tsukamurella	paurometabola	DSM_20162	GCF_000092225.1
Brevibacterium	flavum	ZL-1	GCA_001683055.1

Table A2: Actinobacteria species that were a part of the 252 dataset.

252 Actinbacteria Dataset			
Genus	**Species**	**Strain**	**GCF #**
Actinoalloteichus	fjordicus	ADI127-7	GCF_001941625.1
Actinoalloteichus	hoggarensis	DSM 45943	GCF_002234535.1
Actinoalloteichus	hymeniacidonis	HPA177(T) (=DSM 45092(T))	GCF_001747425.1
Actinopolyspora	erythraea	YIM 90600	GCF_002263515.1
Actinosynnema	mirum	DSM 43827	GCF_000023245.1
Actinosynnema	pretiosum	X47	GCF_002354875.1
Allosaccharopolyspora	coralli	E2A	GCF_009664835.1
Amycolatopsis	acidiphila	KCTC 39523	GCF_021391495.1
Amycolatopsis	aidingensis	YIM 96748	GCF_018885265.1
Amycolatopsis	albispora	WP1	GCF_003312875.1
Amycolatopsis	keratiniphila	HCCB10007	GCF_000400635.2
Amycolatopsis	keratiniphila	MG417-CF17	GCF_000732925.1
Amycolatopsis	mediterranei	RB	GCF_000454025.1
Amycolatopsis	orientalis	B-37	GCF_000943515.2
Corynebacterium	ammoniagenes	MGYG-HGUT-01533	GCF_902381765.1
Corynebacterium	amycolatum	FDAARGOS_1108	GCF_016728725.1
Corynebacterium	anserum	23H37-10	GCF_014262665.1
Corynebacterium	aquilae	S-613	GCF_001941445.1
Corynebacterium	argentoratense	DSM 44202	GCF_000590555.1
Corynebacterium	atypicum	R2070	GCF_000732945.1
Corynebacterium	aurimucosum	DSM 44827; ATCC 700975	GCF_000022905.1
Corynebacterium	callunae	DSM 20147	GCF_000344785.1
Corynebacterium	camporealensis	DSM 44610	GCF_000980815.1
Corynebacterium	casei	LMG S-19264	GCF_000550785.1
Corynebacterium	choanae	200CH	GCF_003813965.1
Corynebacterium	comes	2019	GCF_009734405.1
Corynebacterium	crudilactis	JZ16	GCF_001643015.1
Corynebacterium	cyclohexanicum	ATCC 51369	GCF_020886775.1
Corynebacterium	cystitidis	NCTC11863	GCF_900187295.1
Corynebacterium	deserti	GIMN1.010	GCF_001277995.1
Corynebacterium	diphtheriae	ISS 3319	GCF_002843135.1
Corynebacterium	doosanense	CAU 212	GCF_000767055.1
Corynebacterium	efficiens	YS-314	GCF_000011305.1
Corynebacterium	endometrii	LMM-1653	GCF_004795735.1
Corynebacterium	epidermidicanis	DSM 45586	GCF_001021025.1
Corynebacterium	falsenii	FDAARGOS_1493	GCF_020097615.1

Corynebacterium	flavescens	OJ8	GCF_001941465.1
Corynebacterium	frankenforstense	ST18	GCF_001941485.1
Corynebacterium	gerontici	W8	GCF_003813985.1
Corynebacterium	glaucum	DSM 30827	GCF_002287505.1
Corynebacterium	glucuronolyticum	FDAARGOS_1111	GCF_016728645.1
Corynebacterium	glutamicum	SCgG2	GCF_000404185.1
Corynebacterium	glyciniphilum	AJ 3170	GCF_000626675.1
Corynebacterium	halotolerans	YIM 70093 = DSM 44683	GCF_000341345.1
Corynebacterium	imitans	NCTC13015	GCF_900187215.1
Corynebacterium	jeikeium	K411 = NCTC 11915	GCF_000006605.1
Corynebacterium	kalinowskii	1959	GCF_009734385.1
Corynebacterium	kefirresidentii	FDAARGOS_1055	GCF_016599755.1
Corynebacterium	kroppenstedtii	DSM 44385	GCF_000023145.1
Corynebacterium	kutscheri	DSM 20755	GCF_000980835.1
Corynebacterium	lactis	RW2-5	GCF_001274895.1
Corynebacterium	liangguodongii	2184	GCF_003070865.1
Corynebacterium	lizhenjunii	ZJ-599	GCF_011038655.2
Corynebacterium	lujinxingii	zg-917	GCF_014490555.1
Corynebacterium	marinum	DSM 44953	GCF_000835165.1
Corynebacterium	maris	DSM 45190	GCF_000442645.1
Corynebacterium	minutissimum	NCTC10288	GCF_900478045.1
Corynebacterium	mustelae	DSM 45274	GCF_001020985.1
Corynebacterium	nuruki	S6-4	GCF_007970465.1
Corynebacterium	occultum	2039	GCF_009734425.1
Corynebacterium	pelargi	136/3	GCF_004114895.1
Corynebacterium	phocae	M408/89/1	GCF_001941565.1
Corynebacterium	propinquum	FDAARGOS_1112	GCF_016728665.1
Corynebacterium	pseudopelargi	812CH	GCF_003814005.1
Corynebacterium	pseudotuberculosis	MEX29	GCF_001865765.1
Corynebacterium	qintianiae	MC1420	GCF_011038645.2
Corynebacterium	renale	NCTC7448	GCF_900478035.1
Corynebacterium	resistens	DSM 45100	GCF_000177535.2
Corynebacterium	rouxii	FRC0190	GCF_902702935.1
Corynebacterium	sanguinis	CCUG 58655	GCF_007641235.1
Corynebacterium	silvaticum	PO100/5	GCF_002162115.1
Corynebacterium	simulans	Wattiau	GCF_001586235.1
Corynebacterium	singulare	IBS B52218	GCF_000833575.1
Corynebacterium	sphenisci	DSM 44792	GCF_001941505.1
Corynebacterium	stationis	622=DSM 20302	GCF_001941345.1
Corynebacterium	striatum	FDAARGOS_1115	GCF_016728105.1

Corynebacterium	suranareeae	N24	GCF_002355155.1
Corynebacterium	terpenotabidum	Y-11	GCF_000418365.1
Corynebacterium	testudinoris	DSM 44614	GCF_001021045.1
Corynebacterium	tuberculostearicum	FDAARGOS_1117	GCF_016728365.1
Corynebacterium	uberis	18M0132	GCF_020616335.1
Corynebacterium	ulcerans	5146	GCF_000769635.1
Corynebacterium	urealyticum	NCTC12011	GCF_900187235.1
Corynebacterium	ureicelerivorans	IMMIB RIV-2301	GCF_000747315.1
Corynebacterium	urogenitale	LMM-1652	GCF_009026825.1
Corynebacterium	uterequi	DSM 45634	GCF_001021065.1
Corynebacterium	vitaeruminis	DSM 20294	GCF_000550805.1
Corynebacterium	xerosis	GS 1	GCF_003641245.1
Corynebacterium	yudongzhengii	2183	GCF_003065405.1
Dietzia	lutea	YIM 80766	GCF_003096075.1
Dietzia	psychralcaliphila	ILA-1	GCF_003096095.1
Dietzia	timorensis	ID05-A0528	GCF_001659785.1
Gordonia	alkanivorans	YC-RL2	GCF_004011905.1
Gordonia	amarae	BEN372	GCF_009914515.1
Gordonia	amicalis	CEGA1	GCF_022059885.1
Gordonia	bronchialis	DSM 43247	GCF_000024785.1
Gordonia	hongkongensis	JCM 31934	GCF_023078355.1
Gordonia	insulae	MMS17-SY073	GCF_003855095.1
Gordonia	iterans	Co17	GCF_002993285.1
Gordonia	jinghuaiqii	zg-686	GCF_014041935.1
Gordonia	otitidis	FDAARGOS_1600	GCF_020735545.1
Gordonia	phthalatica	QH-11	GCF_001305675.1
Gordonia	polyisoprenivorans	VH2	GCF_000247715.1
Gordonia	pseudoamarae	BEN371	GCF_009914535.1
Gordonia	rubripertincta	SD5	GCF_014070455.1
Gordonia	zhaorongruii	HY186	GCF_007559005.1
Hoyosella	subflava	DQS3-9A1	GCF_000214175.1
Kibdelosporangium	phytohabitans	KLBMP1111	GCF_001302585.1
Kutzneria	albida	DSM 43870	GCF_000525635.1
Lawsonella	clevelandensis	X1036	GCF_001293125.1
Lentzea	guizhouensis	DHS C013	GCF_001701025.1
Mycobacterium	avium	M011	GCF_016756055.1
Mycobacterium	basiliense	901379	GCF_900292015.1
Mycobacterium	branderi	JCM 12687	GCF_010728725.1
Mycobacterium	canettii	CIPT 140010059	GCF_000253375.1
Mycobacterium	conspicuum	JCM 14738	GCF_010730195.1

Mycobacterium	cookii	JCM 12404	GCF_010727945.1
Mycobacterium	diernhoferi	ATCC 19340	GCF_019456655.1
Mycobacterium	dioxanotrophicus	PH-06	GCF_002157835.1
Mycobacterium	doricum	JCM 12405	GCF_010728155.1
Mycobacterium	florentinum	JCM 14740	GCF_010730355.1
Mycobacterium	grossiae	DSM 104744	GCF_008329645.1
Mycobacterium	haemophilum	ATCC 29548	GCF_000340435.2
Mycobacterium	heckeshornense	JMUB5695	GCF_016861545.1
Mycobacterium	heidelbergense	JCM 14842	GCF_010730745.1
Mycobacterium	holsaticum	JCM 12374	GCF_019645835.1
Mycobacterium	intracellulare	ATCC 13950	GCF_000277125.1
Mycobacterium	kansasii	ATCC 12478	GCF_000157895.3
Mycobacterium	koreense	JCM 19956	GCF_010731835.1
Mycobacterium	koreense	DSM 45575	GCF_022370835.1
Mycobacterium	lacus	JCM 15657	GCF_010731535.1
Mycobacterium	leprae	MRHRU-235-G	GCF_003253775.1
Mycobacterium	lepromatosis	FJ924	GCF_000975265.2
Mycobacterium	malmoense	ATCC 29571	GCF_019645855.1
Mycobacterium	mantenii	JCM 18113	GCF_010731775.1
Mycobacterium	marinum	MMA1	GCF_016745295.1
Mycobacterium	marseillense	FLAC0026	GCF_002285715.1
Mycobacterium	noviomagense	JCM 16367	GCF_010731635.1
Mycobacterium	ostraviense	FDAARGOS_1613	GCF_021183725.1
Mycobacterium	paragordonae	49061	GCF_003614435.1
Mycobacterium	paraseoulense	JCM 16952	GCF_010731655.1
Mycobacterium	paraterrae	DSM 45127	GCF_022430545.1
Mycobacterium	parmense	JCM 14742	GCF_010730575.1
Mycobacterium	pseudoshottsii	JCM 15466	GCF_003584745.1
Mycobacterium	senegalense	ATCC 35796	GCF_019645875.1
Mycobacterium	seoulense	JCM 16018	GCF_010731595.1
Mycobacterium	shigaense	JCM 32072	GCF_002356315.1
Mycobacterium	shinjukuense	JCM 14233	GCF_010730055.1
Mycobacterium	shottsii	JCM 12657	GCF_010728525.1
Mycobacterium	simiae	JCM 12377	GCF_010727605.1
Mycobacterium	spongiae	FSD4b-SM	GCF_018278905.1
Mycobacterium	stomatepiae	JCM 17783	GCF_010731715.1
Mycobacterium	tuberculosis	H37Rv	GCF_000195955.2
Mycobacterium	ulcerans	JKD8049	GCF_020616615.1
Mycobacterium	vicinigordonae	24	GCF_013466425.1
Mycobacterium	virginiense	DSM 100883	GCF_022374935.1

Mycobacteroides	abscessus	GZ002	GCF_004028015.1
Mycobacteroides	abscessus	JCM 16370	GCF_019456675.1
Mycobacteroides	chelonae	CCUG 47445	GCF_001632805.1
Mycobacteroides	chelonae	NJB0901	GCF_002356335.1
Mycobacteroides	immunogenum	CCUG 47286	GCF_001605725.1
Mycobacteroides	salmoniphilum	DSM 43276	GCF_004924335.1
Mycobacteroides	saopaulense	EPM10906	GCF_001456355.1
Mycolicibacter	hiberniae	JCM 13571	GCF_010729485.1
Mycolicibacter	minnesotensis	JCM 17932	GCF_010731755.1
Mycolicibacter	sinensis	JDM601	GCF_000214155.1
Mycolicibacter	sinensis	JCM 6391	GCF_010726505.1
Mycolicibacter	terrae	NCTC10856	GCF_900187145.1
Mycolicibacterium	aichiense	JCM 6376	GCF_010726245.1
Mycolicibacterium	aichiense	JCM 16369	GCF_022370635.1
Mycolicibacterium	alvei	JCM 12272	GCF_010727325.1
Mycolicibacterium	anyangense	JCM 30275	GCF_010731855.1
Mycolicibacterium	arabiense	JCM 18538	GCF_010731815.2
Mycolicibacterium	aurum	NCTC10437	GCF_900637195.1
Mycolicibacterium	baixiangningiae	LJ126	GCF_016313185.1
Mycolicibacterium	boenickei	JCM 15653	GCF_010731295.1
Mycolicibacterium	boenickei	BKK/CU-MFGFA-001	GCF_019739075.1
Mycolicibacterium	celeriflavum	JCM 18439	GCF_010731795.1
Mycolicibacterium	chitae	NCTC10485	GCF_900637205.1
Mycolicibacterium	confluentis	JCM 13671	GCF_010729895.1
Mycolicibacterium	duvalii	JCM 6396	GCF_010726645.1
Mycolicibacterium	fallax	JCM 6405	GCF_010726955.1
Mycolicibacterium	fortuitum	CT6	GCF_001307545.1
Mycolicibacterium	gadium	JCM 12688	GCF_010728925.1
Mycolicibacterium	gilvum	Spyr1	GCF_000184435.1
Mycolicibacterium	hassiacum	Mhassiacum	GCF_900603025.1
Mycolicibacterium	helvum	JCM 30396	GCF_010731895.1
Mycolicibacterium	insubricum	JCM 16366	GCF_010731615.1
Mycolicibacterium	litorale	JCM 17423	GCF_010731695.1
Mycolicibacterium	madagascariense	JCM 13574	GCF_010729665.1
Mycolicibacterium	mageritense	JCM 12375	GCF_010727475.1
Mycolicibacterium	mageritense	ATCC 700504	GCF_022370755.1
Mycolicibacterium	mengxianglii	Z-34	GCF_015710575.1
Mycolicibacterium	monacense	JCM 15658	GCF_010731575.1
Mycolicibacterium	monacense	JCM 16372	GCF_022374875.1
Mycolicibacterium	moriokaense	JCM 6375	GCF_010726085.1

Mycolicibacterium	mucogenicum	DSM 44124	GCF_005670685.2
Mycolicibacterium	neoaurum	MN2019	GCF_018363015.1
Mycolicibacterium	parafortuitum	JCM 6367	GCF_010725485.1
Mycolicibacterium	phocaicum	JCM 15301	GCF_010731115.1
Mycolicibacterium	poriferae	JCM 12603	GCF_010728325.1
Mycolicibacterium	psychrotolerans	JCM 13323	GCF_010729305.1
Mycolicibacterium	pulveris	JCM 6370	GCF_010725725.1
Mycolicibacterium	sarraceniae	JCM 30395	GCF_010731875.1
Mycolicibacterium	sediminis	JCM 17899	GCF_010731735.1
Mycolicibacterium	thermoresistibile	NCTC10409	GCF_900187065.1
Mycolicibacterium	tokaiense	JCM 6373	GCF_010725885.1
Mycolicibacterium	vaccae	95051	GCF_001655245.1
Mycolicibacterium	vanbaalenii	PYR-1	GCF_000015305.1
Nakamurella	multipartita	DSM 44233	GCF_000024365.1
Nakamurella	antarctica	S14-144	GCF_003860405.1
Nocardia	asteroides	NCTC11293	GCF_900637185.1
Nocardia	brasiliensis	FDAARGOS_352	GCF_002209125.2
Nocardia	cyriacigeorgica	GUH-2	GCF_000284035.1
Nocardia	farcinica	IFM 10152	GCF_000009805.1
Nocardia	gipuzkoensis	FDAARGOS_1619	GCF_021183645.1
Nocardia	huaxiensis	WCH-YHL-001	GCF_013744875.1
Nocardia	iowensis	NRRL 5646	GCF_019222765.1
Nocardia	mangyaensis	Y48	GCF_001886715.1
Nocardia	otitidiscaviarum	NEB252	GCF_007362295.1
Nocardia	otitidiscaviarum	DSM 44893	GCF_014217765.1
Nocardia	seriolae	UTF1	GCF_002356035.1
Nocardia	seriolae	CFH S0057	GCF_018362975.1
Nocardia	terpenica	NC_YFY_NT001	GCF_002568625.1
Nocardia	wallacei	FMUON74	GCF_014466955.1
Nocardia	yamanashiensis	FDAARGOS_1623	GCF_021183565.1
Prauserella	marina	DSM 45268	GCF_002240355.1
Pseudonocardia	autotrophica	NBRC 12743	GCF_003945385.1
Pseudonocardia	broussonetiae	Gen 01	GCF_013155125.1
Rhodococcus	aetherivorans	CBO21-1	GCF_021165915.1
Rhodococcus	coprophilus	NCTC10994	GCF_900478115.1
Rhodococcus	erythropolis	R138	GCF_000696675.2
Rhodococcus	erythropolis	CS98	GCF_015099595.1
Rhodococcus	fascians	D188	GCF_001620305.1
Rhodococcus	globerulus	D757	GCF_019334125.1
Rhodococcus	hoagii	DSSKP-R-001	GCF_003013675.1

Rhodococcus	koreensis	R85	GCF_017068375.1
Rhodococcus	opacus	PD630	GCF_020542785.1
Rhodococcus	pyridinivorans	SB3094	GCF_000511305.1
Rhodococcus	pyridinivorans	C1	GCF_016804345.1
Saccharomonospora	viridis	DSM 43017	GCF_000023865.1
Saccharopolyspora	erythraea	NRRL 2338	GCF_000062885.1
Saccharopolyspora	gregorii	MLY014	GCF_022828475.1
Saccharopolyspora	pogona	NRRL30141	GCF_014697215.1
Saccharopolyspora	spinosa	CCTCC M206084	GCF_014490055.1
Saccharothrix	espanaensis	type strain: DSM 44229	GCF_000328705.1
Saccharothrix	syringae	NRRL B-16468	GCF_009498035.1
Segniliparus	rotundus	DSM 44985	GCF_000092825.1
Skermania	piniformis	DSM 43998	GCF_019285775.1
Tomitella	fengzijianii	HY188	GCF_007559025.1
Tomitella	gaofuii	HY172	GCF_014126825.1
Tsukamurella	paurometabola	DSM 20162	GCF_000092225.1

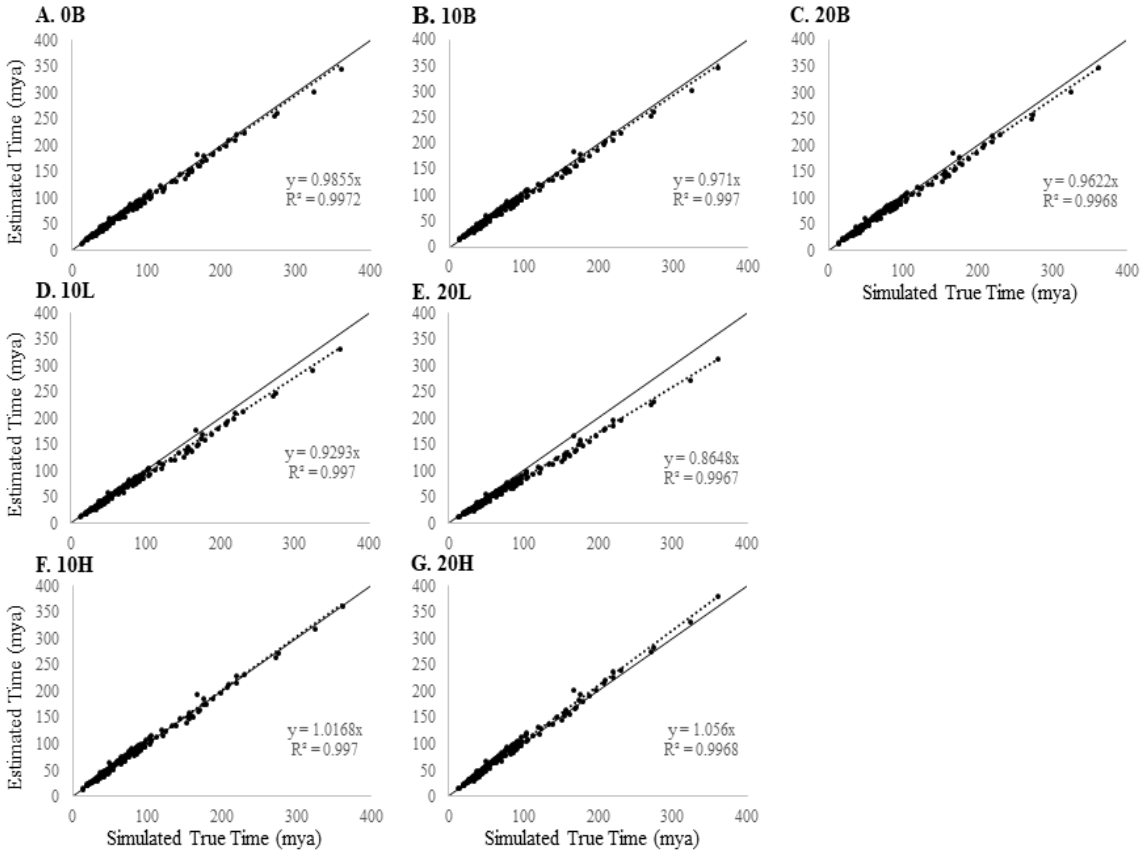


Figure A1. ET accuracy (estimated time *vs.* simulated true time) for tree A with three primary calibrations. The solid line represents a one-to-one match. Equation and R^2 values are shown for each scenario.

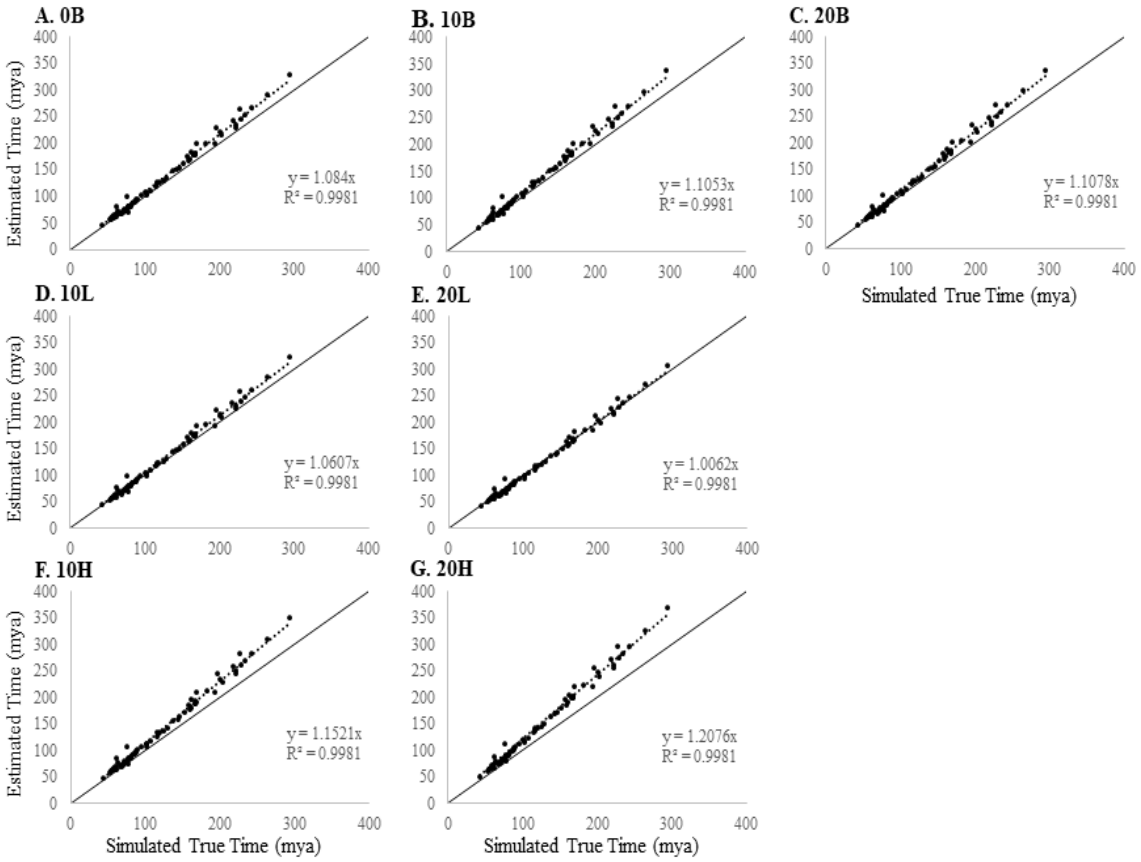


Figure A2. ET accuracy (estimated time *vs.* simulated true time) for tree B with one secondary calibrations. The solid line represents a one-to-one match. Equation and R^2 values are shown for each scenario.

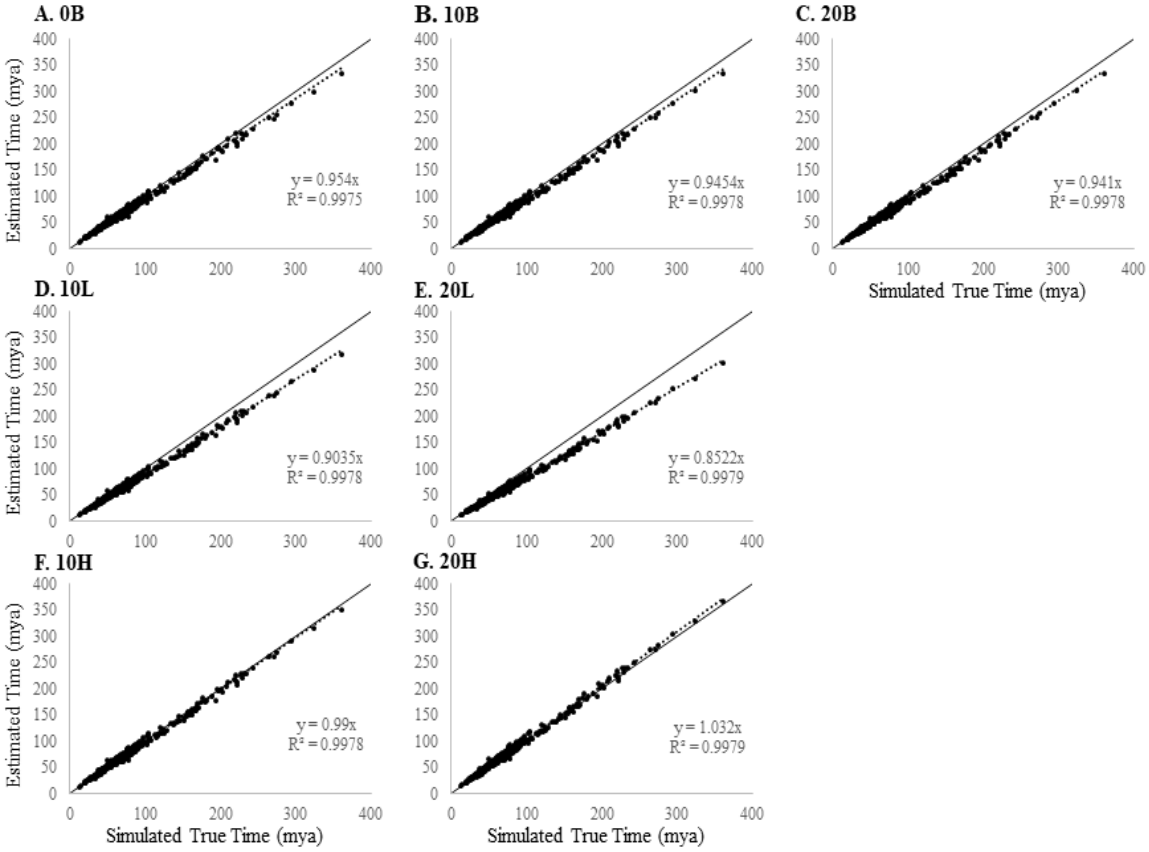


Figure A3. ET accuracy (estimated time *vs.* simulated true time) for tree AB with three distant primary calibrations. The solid line represents a one-to-one match. Equation and R^2 values are shown for each scenario.

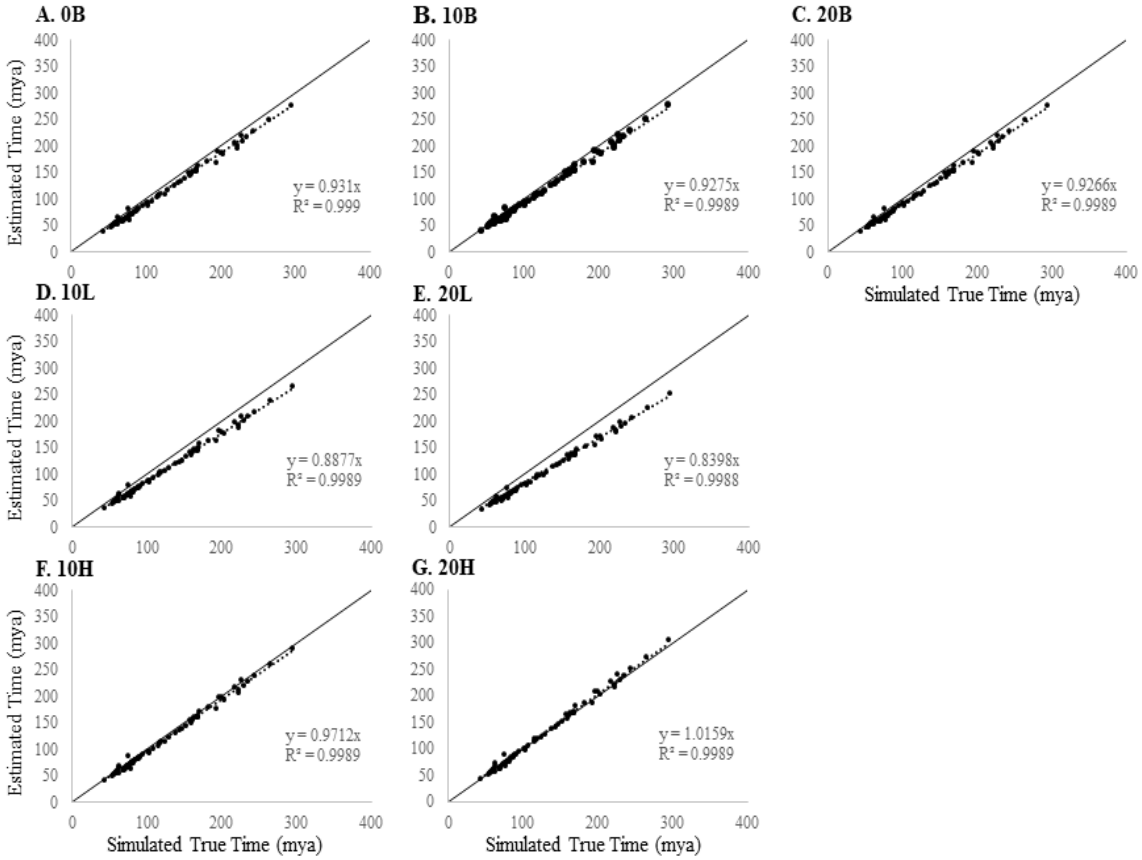


Figure A4. ET accuracy (estimated time *vs.* simulated true time) for tree AB with three distant primary calibrations were only data points for Tree B are plotted. The solid line represents a one-to-one match. Equation and R^2 values are shown for each scenario.

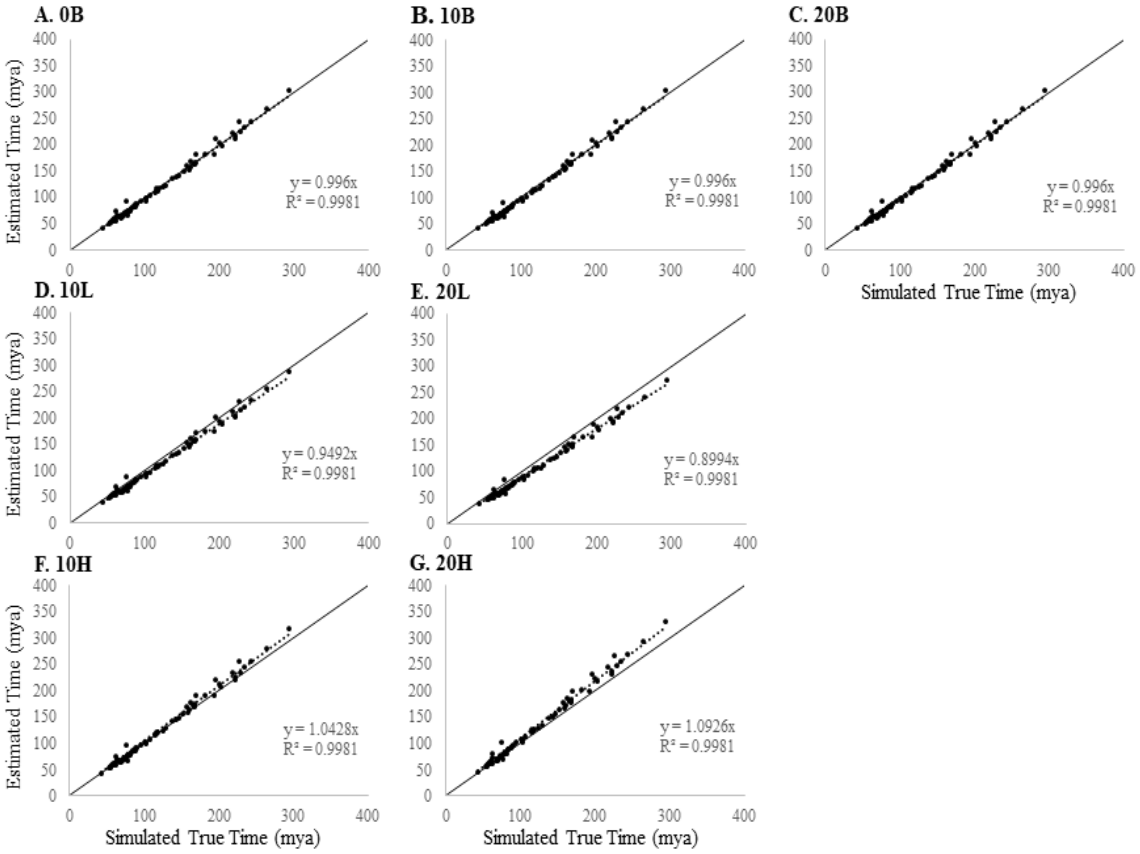


Figure A5. ET accuracy (estimated time *vs.* simulated true time) for Tree B with one primary calibration. The solid line represents a one-to-one match. Equation and R^2 values are shown for each scenario.

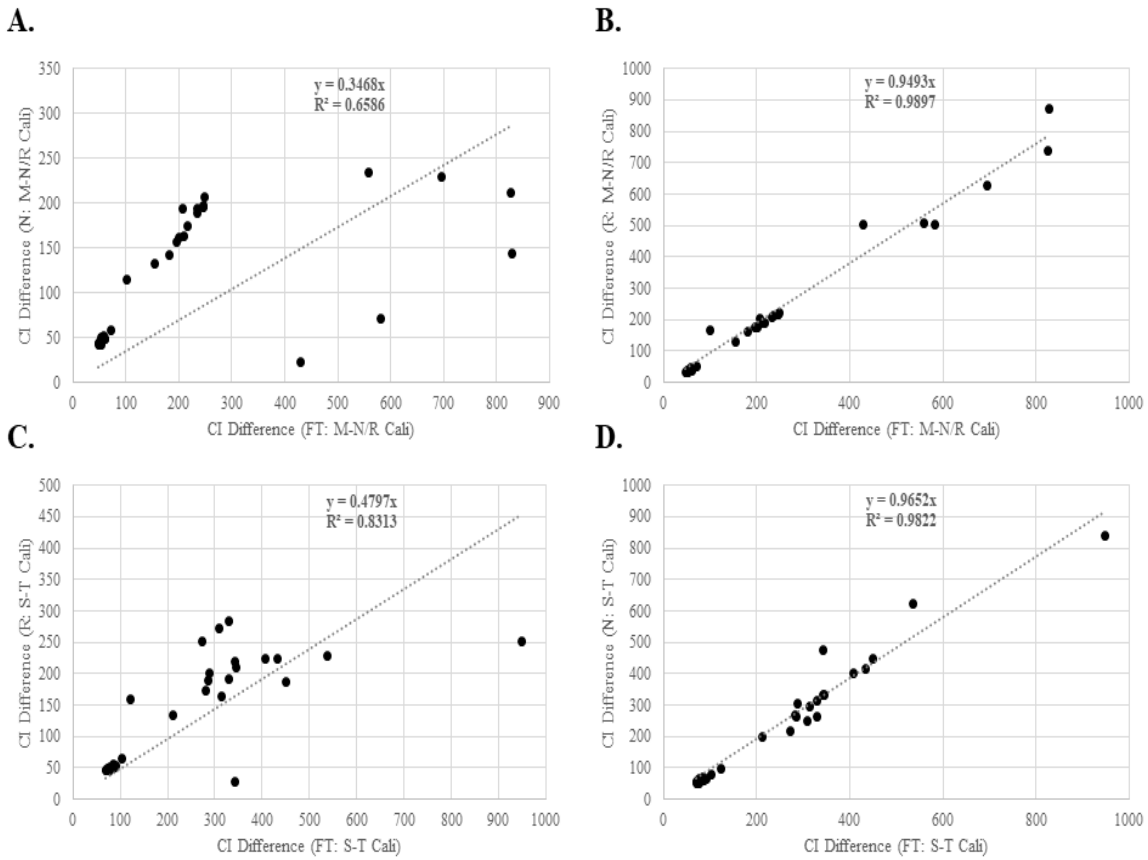


Figure A6: Plotted confidence interval ranges for each node of the KeepOne permutation against the full tree that contains the same calibrations. (**A.**) Nocardia KeepOne permutation vs. full tree calibrated with the Mycobacterium-Nocardia/Rhodococcus node. (**B.**) Rhodococcus KeepOne permutation vs full tree calibrated with the Mycobacterium-Nocardia/Rhodococcus node. (**C.**) Rhodococcus KeepOne permutation vs. full tree calibrated with the Segniliparus – Tsukamurella node. (**D.**) Nocardia KeepOne permutation vs full tree calibrated with the Segniliparus – Tsukamurella node. *CI : confidence interval; N : Nocardia; R : Rhodococcus; M-N/R : Mycobacterium-Nocardia/Rhodococcus; S-T : Segniliparus – Tsukamurella.*