Systematic Bias in Phylogenetic Inference: Implications, Assessment, and Reduction

Suha Naser-Khdour

A thesis submitted for the degree of Doctor of Philosophy of

The Australian National University

February 2022

DECLARATION

I declare that this thesis comprises four original chapters that all are my original research and have been co-authored with my supervisors, Rob Lanfear and Minh Bui. All the chapters of this thesis have been published, submitted or are going to be submitted to peer-reviewed journals as original articles. The text of the published and submitted chapters has not been altered and any structural differences between chapters reflect the requirements of the relevant journal. No part of this has been previously submitted for any other degree.

Suha Naser-Khdour

February 2022

ACKNOWLEDGEMENTS

My first and big appreciation goes to my main supervisor, Assoc. Prof. Robert Lanfear, for his invaluable supervision, guidance and encouragement. Sincere gratitude is extended to his generous participation in guiding, constructive feedback, kind support, and advice during my PhD. Thank you very much, Rob, it was an honour to work with you and learn from you.

I would like to thank my co-supervisors, Dr Minh Bui and Prof. Eric Stone for their continued support, insightful comments and encouragement. My completion of this thesis could not have been accomplished without the support of my team members, Dr Thomas Wong, Caitlin Cherry, Raymond (Weiwen Wang) and Wenqi Zhang. The meetings and conversations were vital in inspiring me to think outside the box. I extend my thanks to everyone at the E&E department for an unforgettable experience.

Last but not least, I would like to thank my family, especially my mother, who set me off on the road to this PhD a long time ago. Finally, my caring, loving, and supportive husband, Nasser; I simply couldn’t have done this without you, my deepest gratitude. Your encouragement and patience when times got rough are much appreciated and duly noted. My accomplishments and success are because you believed in me, and for that, I will be forever in your debt.

LIST OF PUBLICATIONS

Publications included in this thesis:

Naser-Khdour S, Minh BQ, Zhang W, Stone EA, Lanfear R. 2019. The Prevalence and Impact of Model Violations in Phylogenetic Analysis. Genome Biol Evol 11:3341–3352.

Naser-Khdour S, Lanfear R, Minh BQ. 2021. The Influence of Model Violation on Phylogenetic Inference: A Simulation Study. bioRxiv:2021.2009.2022.461455.

Naser-Khdour S, Minh BQ, Lanfear R. 2021. Assessing Confidence in Root Placement on Phylogenies: An Empirical Study Using Non-Reversible Models for Mammals. Syst. Biol.

Publications not included in this thesis:

Ly-Trong N, **Naser-Khdour S**, Lanfear R, Minh BQ. 2021. AliSim: A Fast and Versatile Phylogenetic Sequence Simulator For the Genomic Era. bioRxiv:2021.2012.2016.472905.

ABSTRACT

Molecular phylogenetic inference is the process of reconstructing relationships between individuals, species, or higher groups from genomic sequence data. The reliability of phylogenetic analysis relies on the fit between the substitution models used and the evolutionary processes that generated the data. In phylogenetic inference, we commonly use substitution models which assume that sequence evolution is stationary, reversible, and homogeneous (SRH). Many empirical and simulation studies have shown that assuming SRH conditions can lead to significant errors in phylogenetic inference when the data violates these assumptions. Yet, the extent of SRH violations and their effects on phylogenetic inference of tree topologies are not very well understood.

In Chapter 1, I introduced and applied the Maximal matched-pairs tests of homogeneity (MaxSym tests) to assess the scale and impact of SRH model violations on 3,572 partitions from 35 published phylogenetic data sets. I showed that roughly one-quarter of all the partitions I analysed reject the SRH assumptions and that for more than one-quarter of data sets, tree topologies inferred from all partitions differ significantly from topologies inferred using the subset of partitions that do not reject the SRH assumptions.

In Chapter 2, I simulated datasets under various degrees of non-SRH conditions using empirically derived parameters to mimic real data and examine the effects of incorrectly assuming SRH conditions on inferring phylogenies. I showed that maximum likelihood inference is generally quite robust to a wide range of SRH model violations but is inaccurate under extreme convergent evolution. In addition, I tested the power of the MaxSym tests and other popular tests to detect model violations due to non-SRH evolution. I showed that MaxSym tests performed well under the different schemes of simulations and that of all the

tests I studied, the MaxSym tests perform the best at identifying datasets that might mislead phylogenetic inference.

In Chapter 3, I investigated the homogeneity assumption widely used in phylogenetic inference. To check for homogeneity in empirical datasets, I introduced a computationally feasible test for homogeneity across lineages based on the AIC score. Using empirical datasets from three different clades of life I tested the homogeneity assumption by estimating amino-acid substitution matrices for monophyletic sub-clades within each dataset. I show that forcing the models to be homogenous always provides a worse fit to the data than allowing each sub-clade to have its own model. In addition, for every dataset, I found that a simpler model where two or more clades share the same substitution matrix is always better than the fully non-homogeneous model in terms of AIC score.

In Chapter 4, I investigated the ability of non-reversible models to estimate the root of a phylogeny. In addition, I introduced a new measure of support for the placement of the root in a phylogenetic tree, the *rootstrap* support. I tested the ability of non-reversible models to recover the root placement of five clades of mammals for which prior studies give very strong evidence of a particular root position. I showed that the non-reversible model correctly inferred the root of all the five clades with very high rootstrap support. I then applied the same approaches to infer the roots of two clades of mammals for which previous studies have repeatedly disagreed on the root position. I show that non-reversible models recover similar roots to previous studies, but the rootstrap support is lower than the other five clades.

Together, these chapters show the impact of model violation due to non-SRH evolution on phylogenetic inference and suggest the need to test for model violation prior to phylogenetic inference or to develop and apply more complex substitution models to relax some of the assumptions associated with the most widely used models in phylogenetics.

TABLE OF CONTENTS

INTRODUCTION

Phylogenetics is an essential tool for inferring evolutionary relationships between individuals, species, genes, and genomes (Felsenstein 2004; Bromham 2016; Kapli, et al. 2020). Understanding evolution is vital for our understanding of biology. Traditionally, morphological characters and fossils were used to investigate evolutionary relationships. However, with the advancement and the rapidly dropping costs of DNA sequencing, and the huge progress in computer software and hardware, using molecular data has become the most popular approach in phylogenetic inference.

1.1 The Statistical Model of Substitutions

Based on Darwin's hypothesis of a single origin of life (Darwin 1859), Edwards and Cavalli-Sforza suggested using statistical methods for phylogenetic inference (Edwards and Cavalli-Sforza 1963, 1964). They argued that "probabilistic reasoning leads naturally to the Darwin principle" (Edwards 1996). In a probabilistic framework, we use parametric models of molecular evolution that are designed to approximate the evolutionary process of accumulating changes in the data. Figure 1.1 shows a schematic illustration of the full statistical model used in phylogenetic inference.

$$t,\ \lambda,\mu \xrightarrow{\text{BP}} T' \xrightarrow{\text{S}} T \xrightarrow{\text{M}} \mathbf{X}$$

Figure 1.1| schematic illustration of the model underlying statistical inference of phylogenies. **BP** is the branching process modelling speciation with speciation rate λ and extinction rate μ. Starting from a single origin, and given the elapsed time t, this process will result in a phylogeny T'. **S** is the process of selection by the taxonomist of the samples to be studied among the tips of T', whose phylogeny is T. **M** is the Markov process which operates on the branches of phylogeny T resulting in the observed data X.

The main product of phylogenetic inference is a phylogeny or a phylogenetic tree, which is a graph that connects a group of taxa (external nodes or leaves) with their hypothetical ancestors (internal nodes) by edges (branches). A phylogenetic tree can either be rooted or unrooted and the degree of each node can be determined by the number of branches connected to that node. In phylogenetic inference, we are mainly interested in binary trees (or bifurcating trees) in which all nodes' degrees are not larger than 3 and the root's degree is no larger than 2 (**Figure 1.2**). A node with more than 3 branches connected to it (or root node with more than 2 branches connected to it) represents a polytomy, phylogenetic trees that contain polytomies are called multifurcating trees (**Figure 1.2**). In phylogenetics, a polytomy indicates a lack of information (soft polytomy) or simultaneous divergence event of three or more lineages (hard polytomy) which is highly unlikely.

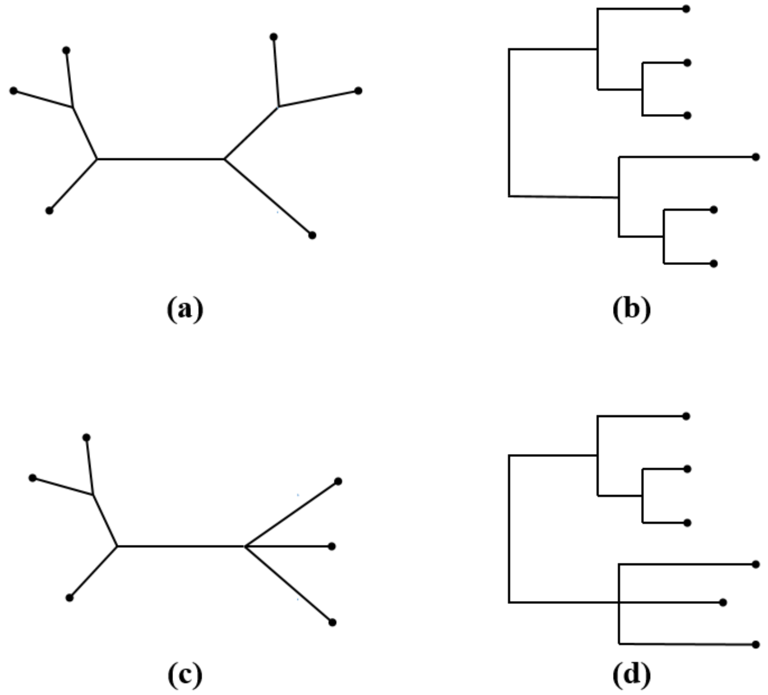


Figure 1.2| Five-taxon **(a)** unrooted bifurcating tree **(b)** rooted bifurcating tree **(c)** unrooted multifurcating tree **(d)** rooted multifurcating tree.

In order to estimate phylogeny T from data X (Figure 1.1) we first need some observed data. Most phylogenetic studies nowadays use DNA or amino acid sequences to infer phylogenies.

DNA and amino acid characters undergo many changes over time and the evolutionary processes underlying these changes are assumed to be **Markovian** and **independent**, implying that changing from one state to another depends only on the current state of the character and all characters evolve independently from other characters. Since the history of substitutions at a site is unknown, yet, the current state is known, it is convenient to assume that the evolutionary process satisfies the Markov property.

For a character *X* in discrete space ({A, C, G, T} for DNA or {A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V} for amino acids), the value of X at time t is denoted by $X(t)$. The probability of the character *X* that currently is in state *i* to be in state *j* after a period time of *t* (*t* > 0) is $P_{ij}(t)$. Since the state space of characters is finite (4 for DNA, 20 for amino acid), the probability distribution can be presented by a transition-probability matrix *P(t)* with elements $P_{ij}(t)$.

$$P_{ij}(t) = P[X(t) = j|X(0) = i], \qquad t > 0 \tag{1.1}$$

$P_{ij}(t)$ is the probability of character X to be in state *j* in time *t* ($P(X(t) = j)$) if it was in state i in time 0 ($P(X(0) = i)$).

Given enough time ($t \to \infty$), the Markov process reaches a stationary distribution π. Once the process reaches its stationary distribution, it will stay in that distribution. Most phylogenetic models assume that the substitution process reached its stationary distribution and therefore all the sequences have the same base composition (Jukes and Cantor 1969; Dayhoff, et al. 1978; Kimura 1980; Felsenstein 1981; Hasegawa, et al. 1985; Tavaré 1986; Tamura and Nei 1993; Whelan and Goldman 2001; Le and Gascuel 2008; Minh, et al. 2021). This assumption is known as the ***stationarity*** assumption and can be represented by equation 1.2:

$$\pi P(t) = \pi, \qquad t > 0 \tag{1.2}$$

Which implies that the base frequencies π remains the same through the time.

In addition, most phylogenetic models assume that the substitution process is ***time-reversible***, meaning that the substitution process remains the same in both directions. It is simple to show that a reversible process is also a stationary process, yet, a stationary process does not have to be reversible (Jermiin, et al. 2017). A time-reversible Markov process should satisfy equation 1.3:

$$\pi_i P_{ij}(t) = P_{ji}\pi_j, \qquad \forall i,j \tag{1.3}$$

Where π_i is the proportion of time the Markov process spends in state *i*, and $\pi_i P_{ij}(t)$ is the amount of flow from state *i* to *j*, while $P_{ji}\pi_j$ is the flow in the opposite direction.

This implies that there exists a distribution π that satisfies equation 1.3 and this distribution is the stationary distribution of the Markov process. In other words, a Markov process that satisfies equation 1.3 is stationary. On the other hand, a Markov process can be stationary and yet its stationary distribution does not satisfy this equation.

Another very common assumption made in phylogenetic inference is that the substitution process is **time-homogeneous** and therefore the instantaneous rate of change from one state to another does not change over time. If the process is assumed to be time-homogeneous, the transition-probability matrix *P(t)* can be found by solving the differential equation:

$$P(t)/dt = P(t)Q, \qquad P(0) = I \tag{1.4}$$

Where *I* is the identity matrix and Q is the instantaneous substitution rate matrix with elements q_{ij}. The solution for this equation is:

$$P(t) = \ e^{Qt} \tag{1.5}$$

For a fully time-homogeneous non-reversible model, the Q matrix will have 12 parameters for DNA (Yang 1994) and 380 parameters for amino acids (Minh, et al. 2020), while in a time-homogeneous, time-reversible model, the Q matrix will have no more than 6 parameters for DNA (Tavaré 1986) and no more than 190 parameters for amino acid (Minh, et al. 2021). The instantaneous rate matrix for DNA sequences with 12 parameters can be represented by:

$$Q = \begin{bmatrix} q_{AA} & q_{AC} & q_{AG} & q_{AT} \\ q_{CA} & q_{CC} & q_{CG} & q_{CT} \\ q_{GA} & q_{GC} & q_{GG} & q_{GT} \\ q_{TA} & q_{TC} & q_{TG} & q_{TT} \end{bmatrix} = \begin{bmatrix} - & \alpha_{AC} & \alpha_{AG} & \alpha_{AT} \\ \alpha_{CA} & - & \alpha_{CG} & \alpha_{CT} \\ \alpha_{GA} & \alpha_{GC} & - & \alpha_{GT} \\ \alpha_{TA} & \alpha_{TC} & \alpha_{TG} & - \end{bmatrix} \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_G & 0 \\ 0 & 0 & 0 & \pi_T \end{bmatrix}$$

Where $q_{ii} = -\sum_{j \neq i} q_{ij}$ and $i, j = \{A, C, G, T\}$, the conditional substitution rate from nucleotide i to j (α_{ij}) is non-negative value and π_j is the frequency of nucleotide j.

From equation 1.3 we can prove that if there exists a stationary distribution π_j, such that

$$\pi_i q_{ij} = \pi_j q_{ji} \tag{1.6}$$

Then:

$$\alpha_{ij} = \alpha_{ji} \tag{1.7}$$

And the model is stationary and time-reversible with just 6 rate parameters. The same is effective for amino-acid sequences, where

$i, j = \{\mathrm{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V}\}$.

Most phylogenetic models are stationary, reversible and homogeneous, these assumptions were made to simplify the mathematical and computational work necessary to infer phylogenies. The main difference between these models is the instantaneous substitution rate matrix (Q). For example the Jukes-Cantor model (Jukes and Cantor 1969) assumes that every nucleotide has the same substitution rate and therefore $q_{ij} = \lambda$ for all $i, j = \{A, C, G, T\}$. Kimura suggested a more complex model that accounts for different transition and transversion rates (Kimura 1980). Both Jukes-Cantor and Kimura models assume that in the stationary distribution the sequence will have equal proprotions of all nucleotides. As this assumption is not realistic, other models such as the TN93 model (Tamura and Nei 1993) and the HKY85 model (Hasegawa, et al. 1985) relax this assumption by allowing unequal base composition in the stationary distribution.

There is no reason to believe that evolution is Stationary, Reversible, or Homogeneous (SRH). In fact, there is huge evidence in the literature that different sequences have different base compositions (Galtier and Gouy 1995; Foster, et al. 1997; Galtier and Gouy 1998; Foster and Hickey 1999; Tarrío, et al. 2001; Paton, et al. 2002; Goremykin and Hellwig 2005; Murray, et al. 2005; Bourlat, et al. 2006; Hyman, et al. 2007; Cox, et al. 2008; Squartini and Arndt 2008; Sheffield, et al. 2009; Nesnidal, et al. 2010; Jayaswal, Jermiin, et al. 2011; Nabholz, et al. 2011; Groussin, et al. 2013; Martijn, et al. 2018; Naser-Khdour, et al. 2019), and therefore the assumption that evolution has reached its stationary distribution is fundamentally wrong. Similarly, there is strong evidence that substitution processes are not reversible (Galtier and Gouy 1998; Galtier, et al. 1999; Jayaswal, et al. 2005; Squartini and Arndt 2008; Woodhams, et al. 2015; Naser-Khdour, et al. 2021) and are not homogeneous (Galtier and Gouy 1998; Galtier, et al. 1999; Herbeck, et al. 2005; Dutheil and Boussau 2008; Jayaswal, Jermiin, et al. 2011; Groussin, et al. 2013).

Initially, the assumptions of stationarity, reversibility, and homogeneity were both necessary to ensure that models of sequence evolution were tractable on early computers, and sensible because early phylogenetic datasets tended to have too little information to estimate phylogenetic trees using highly parameterised models (Kumar, et al. 2012). Various studies have shown that phylogenetic inference can be remarkably robust to violation of these assumptions e.g. (Stiller, et al. 2020; Branstetter, et al. 2021; Maurin, et al. 2021), supporting the often-regurgitated adage that “all models and wrong, but some are useful” (Box 1979). However, technological advances over the last few decades are no longer so limited by computational power or data availability. Furthermore, many of the ‘easy’ branches of the phylogenetic tree of life have been solved, and many modern phylogenetic analyses tend to focus on branches that have been difficult to resolve with certainty using existing methods and models (Rokas and Chatzimanolis 2008). Indeed, many studies have shown that for some branches of the tree of life, small changes to model assumptions or datasets (Delsuc, et al. 2005) can lead to dramatic changes to phylogenetic conclusions. Thus, it is timely to revisit the three fundamental assumptions in phylogenetic analyses, those of stationarity, reversibility, and homogeneity, to ask whether relaxing these assumptions could improve phylogenetic inference

1.2 The Maximum-Likelihood Inference

After obtaining the best-fit Markov model that describes the evolution of the data, there is a need to infer the relationships between the taxa using that model. Most probabilistic phylogenetic methods use maximum likelihood (ML) to infer those relationships in a form of trees (Felsenstein 1981). However, inferring evolutionary trees by ML methods is known to be an NP-hard problem (Chor and Tuller 2005) and therefore many heuristic algorithms are available for ML phylogenetic inference (Swofford 2001; Lemmon and Milinkovitch 2002; Guindon, et al. 2010; Price, et al. 2010; Bazinet, et al. 2014; Stamatakis 2014; Minh, et al. 2020). Those heuristics traverse the tree and model parameter space to find the tree and model with the maximum probability of producing the observed data, yet, they cannot guarantee to find the optimal tree.

A common assumption in ML inference is that different sites evolve independently of each other. Under that assumption, the likelihood of a tree is the product over all sites (Felsenstein 1981). Equivalently, the log-likelihood is the sum over all sites. Equation 1.8 represents the log-likelihood for a dataset with n independent sites

$$\log(L) = \sum_{i=1}^{n} log\{f(x_i|\theta)\} \qquad (1.8)$$

where L is the likelihood of the tree and $f(x_i|\theta)$ is the frequency distribution with unknown parameter θ over site x_i. Then the ML function estimates θ by maximizing the log-likelihood function.

For computational convenience, the vast majority of phylogenetic methods assume stationary, reversible and homogeneous evolution. Besides reducing the dimensionality of the parameter space, assuming time-reversibility, and thus stationarity simplifies the tree search without affecting the likelihood of the inference (Felsenstein 1981). Calculating the likelihood of a tree under a time-reversible model does not require starting from the root and moving forward in time. A time-reversible algorithm can start the computation of the likelihood from any node moving forward or backwards in time as desired until the likelihood of the tree is calculated. This is known as the “Pulley Principle” (Felsenstein 1981) and it allows moving the root of the tree without affecting the likelihood. Moreover, using the pulley principle reduces the tree space since the number of unrooted trees for *n* taxa is always smaller than the number of rooted trees for the same number of taxa (

Table **1**).

Table 1| The number of rooted and unrooted trees for n taxa.

n	No. unrooted trees	No. rooted trees
3	1	3
4	3	15
5	15	105
10	2,027,025	34,459,425
20	2.22×10^{20}	8.20×10^{21}
50	2.84×10^{74}	2.75×10^{76}

1.3 Systematic Bias in Phylogenetic Inference

There are two main types of errors in phylogenetic inference, stochastic (sampling) error and systematic error. Contrary to the stochastic error that can be reduced by adding more characters to the data, increasing the number of characters in a dataset does not seem to reduce systematic bias, and sometimes it might even intensify it (Philippe, et al. 2005; Sullivan and Joyce 2005; Kumar, et al. 2012; Brown and Thomson 2017; Duchene, et al. 2017). The statistical inference has the probability of consistency if it converges toward the true tree as more characters are added to the data (Felsenstein 1978). Thus, the consistency of the ML inference is guaranteed as long as there is no violation of the model's assumptions (Philippe, et al. 2005; Kumar, et al. 2012; Lemmon and Lemmon 2013). Since all phylogenetic inference methods make some assumptions, consistency has become one of the most challenging aspects of phylogenetics. In the era of big datasets and genome-scale sequences, this challenge is more evident now than ever, since it is much easier to get over-confident in the wrong tree (Philippe, et al. 2005; Kumar, et al. 2012).

In a probabilistic framework, inconsistency can be almost always attributed to the systematic error caused by the violation of the model assumptions (Philippe, et al. 2005). The most popular assumptions that are often violated in phylogenetic analysis are rate homogeneity across sites, rate homogeneity across lineages, compositional homogeneity across lineages and reversibility. If the data violate one or more of these assumptions this can lead to convergence towards the wrong tree as more data are added to an analysis, sometimes even with very high statistical support for incorrect inferences (Swofford 2001; Felsenstein and Felenstein 2004; Ho and Jermiin 2004; Jermiin, et al. 2004; Kumar, et al. 2012).

Nowadays, in the era of big datasets and genome-scale sequences, it is common to infer incorrect trees with very high statistical support. Although using big datasets makes the site-sampling variance negligible (Kumar, et al. 2012), it increases the probability of systematic bias. The more characters that are included in datasets, the higher the chance that those characters did not evolve under homogeneous conditions. Generally, small deviations and high confidence in the results are desirable features of phylogenetic inference. However, if the inference suffers a lack of consistency, this feature becomes very misleading. In fact, a growing body of studies presents high confidence in contradicting results (Foster and Hickey 1999; Tarrío, et al. 2001; Paton, et al. 2002; Goremykin and Hellwig 2005; Murray, et al. 2005; Bourlat, et al. 2006; Hyman, et al. 2007; Sheffield, et al. 2009; Nesnidal, et al. 2010; Nabholz, et al. 2011; Martijn, et al. 2018).

There are two main approaches to deal with systematic bias in phylogenetic inference:

I) using complex models with a large number of free parameters: a number of models that relax the one or more of the popular assumptions in phylogenetic inference are available (Foster 2004; Lartillot and Philippe 2004; Blanquart and Lartillot 2006; Boussau and Gouy 2006; Knight, et al. 2007; Dutheil and Boussau 2008; Sumner, et al. 2012; Zou, et al. 2012; Groussin, et al. 2013; Jayaswal, et al. 2014). Yet, they remain relatively rarely used, as searching for optimal phylogenetic trees under these models is computationally demanding even with modern computational resources (Betancur-r, et al. 2013) and the implementations are often not easy to use. As a result, the vast majority of empirical phylogenetic inferences rely on models which assume that sequences have evolved under SRH conditions, such as the general time-reversible (GTR) family of models implemented in the most widely-used phylogenetics software packages (Swofford 2001; Drummond and Rambaut 2007; Guindon, et al. 2010; Ronquist, et al. 2012; Bazinet, et al. 2014; Bouckaert, et al. 2014; Stamatakis 2014; Höhna, et al. 2016; Minh, et al. 2020).

II) testing for model violation in the data: statistical tests for model violations can be applied on the data *a priori* to the phylogenetic inference, and trees can then be reconstructed exclusively from data that do not violate the models. A number of methods have been proposed to test for violation of SRH conditions in aligned sequences prior to estimating trees (Bowker 1948; Stuart 1955; Rzhetsky and Nei 1995; Kumar and Gadagkar 2001; Weiss and von Haeseler 2003; Ababneh, et al. 2006; Ho, et al. 2006), and there are also *a posteriori* tests for absolute model adequacy which are employed after trees have been estimated (Goldman 1993; Foster 2004; Brown and ElDabaje 2009; Brown 2014; Duchene, et al. 2017; Brown and Thomson 2018). However, testing for model violation either pre- or post-analysis also remains relatively rare in the empirical phylogenetic literature.

1.4 Assessing Model Assumptions in Phylogenetic Inference

Validating that the data complies with the assumptions of the model is a key to reducing systematic bias in phylogenetic inference (Philippe, et al. 2005; Brown 2014; Jermiin, Catullo, et al. 2020). As discussed above, the need for methods that assess the evolutionary process prior to phylogenetic inference becomes more important as the number of sequences and sites per dataset increases (Ho and Jermiin 2004; Jermiin, et al. 2004; Phillips, et al. 2004; Delsuc, et al. 2005; Lemmon and Lemmon 2013). In the phylogenetic protocol proposed by Jermiin et al. (Jermiin, Catullo, et al. 2020) the authors suggest two new novel steps to the current phylogenetic protocol used by most researchers: 1) assessing phylogenetic assumptions *a priori* to the phylogenetic inference and 2) testing for goodness-of-fit *a posteriori* to the phylogenetic inference (Figure 1.3).

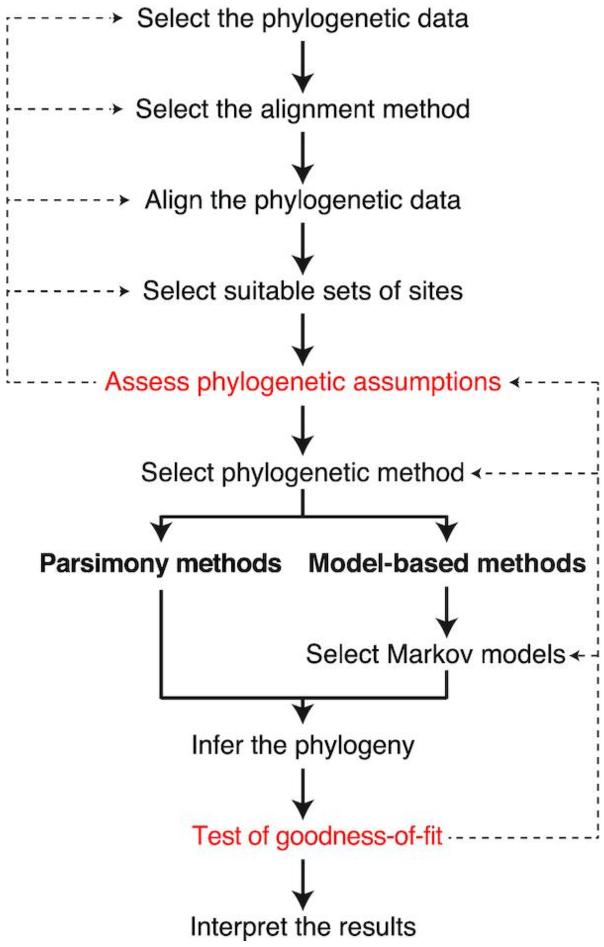

Select the phylogenetic data
Select the alignment method
Align the phylogenetic data
Select suitable sets of sites
Assess phylogenetic assumptions
Select phylogenetic method
Parsimony methods
Model-based methods
Select Markov models
Infer the phylogeny
Test of goodness-of-fit
Interpret the results

Figure 1.3| schematic illustration of the protocol proposed by Jermiin, Catullo, et al. 2020. Solid arrows show the order of actions normally taken during phylogenetic analysis. Dashed arrows show feedback loops often employed in phylogenetic research. This image can be reused under a CC-BY-NC-4.0 license.

Allowing the data to reject the model when its assumptions are violated (step 5 in Fig. 1.3) will alert us to choose more complex models for phylogeny reconstruction or potentially omit these loci from downstream analyses (Kumar and Gadagkar 2001; Jermiin, Catullo, et al. 2020). There are several available methods for assessing SRH assumptions a priori to phylogenetic inference (Kelly 1994; Rzhetsky and Nei 1995; Kumar and Gadagkar 2001; Weiss and von

Haeseler 2003; Ababneh, et al. 2006; Ho, et al. 2006; Squartini and Arndt 2008; Kedzierska, et al. 2012; Jermiin, et al. 2017; Naser-Khdour, et al. 2019; Jermiin, Lovell, et al. 2020) validating that the data complies with the assumptions of the model in use. Although there are other popular assumptions in phylogenetic inference such as; heterotachy (Lopez, et al. 2002), tree-likeness and that the evolutionary process is independent and identically distributed (iid), in this thesis I will only focus on the SRH assumptions.

In addition to assessing the phylogenetic assumptions, it is important to assess if the data can be properly explained by the inferred phylogeny (step 10, Fig. 1.3). Various methods are available to assess the adequacy of the combined model and tree produced by a phylogenetic inference (Goldman 1993; Steel, et al. 1993; Bollback 2002; Foster 2004; Brown and ElDabaje 2009; Brown 2014; Duchene, et al. 2017; Brown and Thomson 2018; Naser-Khdour, et al. 2021). However, to increase the goodness-of-fit it is still important to test for violations of the model assumptions before the phylogenetic analysis. This will allow us to use better data (by removing the parts that violate the assumptions) or better models (that relax the violated assumptions) before the phylogenetic analysis and therefore improve our chances of inferring an accurate phylogeny.

1.5 Complex Models of Evolution

Another approach to account for model violation is to use substitution models that relax some of the most violated assumptions in phylogenetic inference. As discussed in sections 1.1 and 1.2, using substitution models that assume SRH evolution is merely a computational convenience and relaxing one or more of these assumptions will put in more computational burden on the analysis. Moreover, most of these methods are not feasible for large datasets and can only be applied to a small number of taxa and sites.

Although several substitution models that relax one or more of the SRH assumptions are available (Barry and Hartigan 1987; Reeves 1992; Lake 1994; Lockhart, et al. 1994; Galtier and Gouy 1995; Yang and Roberts 1995; Galtier and Gouy 1998; Gu and Li 1998; Galtier, et al. 1999; Tamura and Kumar 2002; Foster 2004; Lartillot and Philippe 2004; Jayaswal, et al. 2005; Blanquart and Lartillot 2006; Boussau and Gouy 2006; Jayaswal, et al. 2007; Blanquart and Lartillot 2008; Dutheil and Boussau 2008; Jayaswal, Ababneh, et al. 2011; Jayaswal, Jermiin, et al. 2011; Dutheil, et al. 2012; Sumner, et al. 2012; Zou, et al. 2012; Groussin, et al. 2013; Jayaswal, et al. 2014; Woodhams, et al. 2015; Minh, et al. 2020), most of these models requires a phylogeny as their input, and then estimate the best-fit non-SRH model for that given phylogeny. Most implementations remain computationally intractable for the task of searching for the ML tree *and* the ML model in combination. Indeed, since phylogeny is usually unknown and is the *target* of most phylogenetic analyses, most of these methods remain rarely used in phylogenetic inference.

1.6 Motivation and aims of the thesis

In this thesis, I aim to reduce the systematic bias associated with model violation by developing and applying new methods that test for model violation *a priori* and *a posteriori* to the phylogenetic analysis. In addition, I will evaluate the potential impact of different types of model violations by comparing the adequacy of SRH and non-SRH models on large collections of empirical datasets.

In Chapter 1, I extend the matched-pair test of symmetry (Bowker 1948), the matched-pair test of marginal symmetry (Stuart 1955) and the matched-pair test of internal symmetry (Ababneh, et al. 2006) to accommodate more than a pair of sequences. These three tests are designed to check for symmetry between two homologous sequences without previous knowledge or assumptions regarding their topology or the evolutionary processes operating on them. Thus these tests are considered as tests for model violation *a priori* to phylogenetic inference. Yet, these tests are designed for pairs of sequences only. Since most, if not all datasets contain more than two sequences, it is vital to propose new tests for multiple-sequence alignments, allowing empirical phylogeneticists to ask whether any individual alignment shows evidence of violating the SRH assumptions.

In addition, using the newly proposed tests with various published empirical datasets from different clades of life, different types of genomes and a varying number of taxa and sites I assess the scale and impact of SRH model violations on phylogenetic inference. Even though there is strong evidence that the SRH assumptions are repeatedly violated by real data (Foster and Hickey 1999; Tarrío, et al. 2001; Paton, et al. 2002; Goremykin and Hellwig 2005; Murray, et al. 2005; Bourlat, et al. 2006; Hyman, et al. 2007; Sheffield, et al. 2009; Nesnidal, et al. 2010; Nabholz, et al. 2011; Martijn, et al. 2018), there is still little known about the prevalence of SRH assumptions' violation in empirical data and its effect on phylogenetic inference. Thus, it is crucial first to understand this phenomenon when addressing the issue of systematic bias due to SRH model violation. We need to know not only which assumptions are violated by empirical datasets, but also which violations have important impacts on downstream inferences. The results of this chapter show that violation of SRH conditions is prevalent across all different types of datasets and that it has a substantial influence on the tree topology.

In Chapter 2, I use simulation to examine the effects of incorrectly assuming SRH conditions on inferring phylogenies. To do this, I simulate thousands of alignments with a various number of taxa, sites, models, and degrees of non-SRH conditions. Although several studies used simulated data to answer this question (Huelsenbeck and Hillis 1993; Hillis, et al. 1994; Galtier and Gouy 1998; Ho and Jermiin 2004; Jermiin, et al. 2004; Boussau and Gouy 2006), the majority of these studies used simulations that reflect extreme cases of convergent evolution, which are unlikely to represent the majority of empirical datasets. Therefore, in this chapter, I derive the parameters for the simulations from tens of thousands of published empirical datasets in order to mimic as closely as possible the evolution of a broad range of published datasets. I then combine these with a new simulation scheme in which model parameters are inherited with modification along a simulated phylogenetic tree. Moreover, I examine the power of the tests that I proposed in the previous chapter, along with a number of other existing tests, to detect model violations due to non-SRH evolution. Since the data is simulated under various degrees of non-SRH evolution, I can estimate type I and type II errors for these tests. The results of this chapter show that maximum likelihood inference is generally quite robust to a wide range of SRH model violations but is inaccurate under extreme convergent evolution. Moreover, I show that the tests I introduced in the previous chapter (namely, the MaxSym tests) are successfully able to detect SRH violations in the simulated alignments and even predict the accuracy of the tree inference.

In Chapter 3, I investigate the homogeneity assumption widely used in phylogenetic inference. The homogeneity assumption implies that the instantaneous substitution rate matrix is constant over the tree (Jermiin, et al. 2017). Therefore, relaxing the homogeneity assumption requires assigning different matrices across the branches of the tree. There have been several attempts to use non-homogeneous matrices for phylogenetic inference (Barry and Hartigan 1987; Roberts and Yang 1995; Galtier and Gouy 1998; Galtier, et al. 1999; Foster 2004; Jayaswal, et al. 2005; Blanquart and Lartillot 2006; Jayaswal, et al. 2007; Blanquart and Lartillot 2008; Dutheil and Boussau 2008; Jayaswal, Jermiin, et al. 2011; Dutheil, et al. 2012; Zou, et al. 2012; Groussin, et al. 2013; Jayaswal, et al. 2014), however, these models did not gain popularity due to computational limitations or due to the requirement for assuming the topology in advance. As a consequence of the lack of easily-applied non-homogeneous models, the true prevalence and effect of homogeneity assumption on phylogenetic inference are not well understood. In this chapter, I will introduce a new algorithm to check for non-homogeneity in large empirical datasets, and apply it to large datasets from three different clades of life; birds (Jarvis, et al. 2015), plants (Ran, et al. 2018), and mammals (Wu, et al. 2018). The results of this chapter reveal that the homogeneous model with one matrix for the whole dataset is significantly worse than non-homogeneous models that use more than one matrix. Moreover, the results show that the fully non-homogeneous model with the maximum possible number of matrices is always worse than a simpler non-homogeneous model with a fewer number of matrices.

In Chapter 4, I assess the ability of non-reversible models to accurately root phylogenetic trees without the need for external information (e.g. outgroup taxa) or other assumptions (e.g. molecular clocks). For that purpose, I use non-reversible substitution models for DNA and amino acid sequences (Minh, et al. 2020) to infer rooted phylogenies of five clades of mammals, (Afrotheria, Bovidae, Carnivora, Primates, and Myomorpha) for which there is strong agreement in their root placement from a large range of previous studies. In addition, I infer the rooted phylogeny for another two clades of mammals, (Chiroptera and Cetartiodactyla) for which there is no consensus regarding their root placements and evaluate the precision of the non-reversible models to estimate their root compared to the other five clades. I also introduce three new metrics to help researchers assess the statistical support for different root placements during phylogenetic analyses regardless of the rooting method. Additionally, I use the AU test (Shimodaira 2002) to provide additional information about the confidence that we have in a certain root placement. The results of this chapter show that non-reversible models can infer the root of the phylogeny with very high accuracy. In addition, the results show that removing loci that fail the MaxSym test improves the support for the correct root placement.

In order to compare the goodness-of-fit of the non-reversible and reversible models to the data, I will use the BIC score (Schwarz 1978). I expect that non-reversible models will have a better fit for the data since they relax one of the major non-realistic assumptions in phylogenetic inference, namely, time-reversibility. Yet, the BIC criterion sometimes can be biased (Susko and Roger 2020) and therefore it is not sufficient to rely on the BIC score (or equivalently on the AIC (Akaike 1974) score) to determine the efficiency of the non-reversible models in inferring the root of a phylogeny. Thus, it is important to have a valid measure for the robustness of the root placement given the model and the data too.

In this chapter, I propose new metrics to assess the extent of statistical support that the data have for a certain root placement: the rootstrap, and an application of the AU test (Shimodaira 2002). The rootstrap describes for each branch in a tree the proportion of bootstrap samples in which that branch was selected as the root branch. And I use the AU test to generate a confidence set of root branches on a single phylogenetic tree. These metrics are not exclusive to non-reversible models and can be used in any rooted tree, regardless of the rooting method.

References

Ababneh F, Jermiin LS, Ma C, Robinson J. 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics 22:1225-1231.

Akaike H. 1974. A new look at the statistical model identification. IEEE transactions on automatic control 19:716-723.

Barry D, Hartigan JA. 1987. Statistical Analysis of Hominoid Molecular Evolution. Statistical Science 2:191-207.

Bazinet AL, Zwickl DJ, Cummings MP. 2014. A gateway for phylogenetic analysis powered by grid computing featuring GARLI 2.0. Syst. Biol. 63:812-818.

Betancur-r R, Li C, Munroe TA, Ballesteros JA, Ortí G. 2013. Addressing gene tree discordance and non-stationarity to resolve a multi-locus phylogeny of the flatfishes (Teleostei: Pleuronectiformes). Syst. Biol. 62:763-785.

Blanquart S, Lartillot N. 2006. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Mol. Biol. Evol. 23:2058-2071.

Blanquart S, Lartillot N. 2008. A site- and time-heterogeneous model of amino acid replacement. Mol. Biol. Evol. 25:842-858.

Bollback JP. 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. 19:1171-1180.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comp. Biol. 10:e1003537.

Bourlat SJ, Juliusdottir T, Lowe CJ, Freeman R, Aronowicz J, Kirschner M, Lander ES, Thorndyke M, Nakano H, Kohn AB. 2006. Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. Nature 444:85.

Boussau B, Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. Syst. Biol. 55:756-768.

Bowker AH. 1948. A test for symmetry in contingency tables. J Am Stat Assoc 43:572-574.

Box GEP. 1979. Robustness in the Strategy of Scientific Model Building. In. Robustness in Statistics. p. 201-236.

Branstetter MG, Müller A, Griswold TL, Orr MC, Zhu CD. 2021. Ultraconserved element phylogenomics and biogeography of the agriculturally important mason bee subgenus Osmia (Osmia). Syst. Entomol. 46:453-472.

Bromham L. 2016. An introduction to molecular evolution and phylogenetics. Oxford, UK: Oxford University Press.

Brown JM. 2014. Detection of Implausible Phylogenetic Inferences Using Posterior Predictive Assessment of Model Fit. Syst. Biol. 63:334-348.

Brown JM, ElDabaje R. 2009. PuMA: Bayesian analysis of partitioned (and unpartitioned) model adequacy. Bioinformatics 25:537-538.

Brown JM, Thomson RC. 2017. Bayes Factors Unmask Highly Variable Information Content, Bias, and Extreme Influence in Phylogenomic Analyses. Syst. Biol. 66:517-530.

Brown JM, Thomson RC. 2018. Evaluating Model Performance in Evolutionary Biology. Annu Rev Ecol Evol S 49:null.

Chor B, Tuller T editors. Annual International Conference on Research in Computational Molecular Biology. 2005.

Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. 2008. The archaebacterial origin of eukaryotes. Proc Natl Acad Sci U S A 105:20356-20361.

Darwin C. 1859. The origin of species by means of natural selection, or the preservation of favored races in the struggle for life. London: Murray.

Dayhoff M, Schwartz R, Orcutt B. 1978. A model of evolutionary change in proteins. Atlas of protein sequence and structure 5:345-352.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics 6:361.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

Duchene DA, Duchene S, Ho SYW. 2017. New Statistical Criteria Detect Phylogenetic Bias Caused by Compositional Heterogeneity. Mol. Biol. Evol. 34:1529-1534.

Dutheil J, Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. BMC Evol. Biol. 8:255.

Dutheil JY, Galtier N, Romiguier J, Douzery EJ, Ranwez V, Boussau B. 2012. Efficient selection of branch-specific models of sequence evolution. Mol. Biol. Evol. 29:1861-1874.

Edwards AW. 1996. The origin and early development of the method of minimum evolution for the reconstruction of phylogenetic trees. Syst. Biol. 45:79-91.

Edwards AWF, Cavalli-Sforza LL editors. Heredity. 1963.

Edwards AWF, Cavalli-Sforza LL. 1964. Reconstruction of evolutionary trees. In: Heywood VH, McNeill J, editors. Phenetic and Phylogenetic Classification. London: Systematic Association. p. 67-76.

Felsenstein J. 1978. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. Syst. Biol. 27:401-410.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368-376.

Felsenstein J. 2004. Inferring Phylogenies. Sunderland, Massachusetts: Sinauer Associates, Inc.

Felsenstein J, Felenstein J. 2004. Inferring phylogenies: Sinauer associates Sunderland, MA.

Foster PG. 2004. Modeling compositional heterogeneity. Syst. Biol. 53:485-495.

Foster PG, Hickey DA. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J. Mol. Evol. 48:284-290.

Foster PG, Jermiin LS, Hickey DA. 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J. Mol. Evol. 44:282-288.

Galtier N, Gouy M. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. 15:871-879.

Galtier N, Gouy M. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proc Natl Acad Sci U S A 92:11317-11321.

Galtier N, Tourasse N, Gouy M. 1999. A nonhyperthermophilic common ancestor to extant life forms. Science 283:220-221.

Goldman N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182-198.

Goremykin V, Hellwig F. 2005. Evidence for the most basal split in land plants dividing bryophyte and tracheophyte lineages. Plant Syst. Evol. 254:93-103.

Groussin M, Boussau B, Gouy M. 2013. A branch-heterogeneous model of protein evolution for efficient inference of ancestral sequences. Syst. Biol. 62:523-538.

Gu X, Li WH. 1998. Estimation of evolutionary distances under stationary and nonstationary models of nucleotide substitution. Proc Natl Acad Sci U S A 95:5899-5905.

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

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160-174.

Herbeck JT, Degnan PH, Wernegreen JJ. 2005. Nonhomogeneous model of sequence evolution indicates independent origins of primary endosymbionts within the enterobacteriales (gamma-Proteobacteria). Mol. Biol. Evol. 22:520-532.

Hillis DM, Huelsenbeck JP, Cunningham CW. 1994. Application and accuracy of molecular phylogenies. Science 264:671-677.

Ho JW, Adams CE, Lew JB, Matthews TJ, Ng CC, Shahabi-Sirjani A, Tan LH, Zhao Y, Easteal S, Wilson SR, et al. 2006. SeqVis: visualization of compositional heterogeneity in large alignments of nucleotides. Bioinformatics 22:2162-2163.

Ho SY, Jermiin L. 2004. Tracing the decay of the historical signal in biological sequence data. Syst. Biol. 53:623-637.

Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, Huelsenbeck JP, Ronquist F. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst. Biol. 65:726-736.

Huelsenbeck JP, Hillis DM. 1993. Success of Phylogenetic Methods in the Four-Taxon Case. Syst. Biol. 42:247-264.

Hyman IT, Ho SY, Jermiin LS. 2007. Molecular phylogeny of Australian Helicarionidae, Euconulidae and related groups (Gastropoda: Pulmonata: Stylommatophora) based on mitochondrial DNA. Mol. Phylogen. Evol. 45:792-812.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT, et al. 2015. Phylogenomic analyses data of the avian phylogenomics project. Gigascience 4:4.

Jayaswal V, Ababneh F, Jermiin LS, Robinson J. 2011. Reducing model complexity of the general Markov model of evolution. Mol. Biol. Evol. 28:3045-3059.

Jayaswal V, Jermiin LS, Poladian L, Robinson J. 2011. Two stationary nonhomogeneous Markov models of nucleotide sequence evolution. Syst. Biol. 60:74-86.

Jayaswal V, Jermiin LS, Robinson J. 2005. Estimation of Phylogeny Using a General Markov Model. Evol Bioinform 1:62-80.

Jayaswal V, Robinson J, Jermiin L. 2007. Estimation of phylogeny and invariant sites under the general Markov model of nucleotide sequence evolution. Syst. Biol. 56:155-162.

Jayaswal V, Wong TK, Robinson J, Poladian L, Jermiin LS. 2014. Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. Syst. Biol. 63:726-742.

Jermiin L, Ho SY, Ababneh F, Robinson J, Larkum AW. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst. Biol. 53:638-643.

Jermiin LS, Catullo RA, Holland BR. 2020. A new phylogenetic protocol: dealing with model misspecification and confirmation bias in molecular phylogenetics. NAR Genom Bioinform 2:lqaa041.

Jermiin LS, Jayaswal V, Ababneh FM, Robinson J. 2017. Identifying Optimal Models of Evolution. In: Keith JM, editor. Bioinformatics. Melbourne: Humana Press, New York, NY. p. 379-420.

Jermiin LS, Lovell DR, Misof B, Foster PG, Robinson J. 2020. Detecting and visualising the impact of heterogeneous evolutionary processes on phylogenetic estimates. bioRxiv:828996.

Jukes TH, Cantor C. 1969. Evolution of protein molecules. In: Munro HN, editor. In Mammalian Protein Metabolism: Academic Press, New York. p. 21–132.

Kapli P, Yang Z, Telford MJ. 2020. Phylogenetic tree building in the genomic age. Nat. Rev. Genet. 21:428-444.

Kedzierska AM, Drton M, Guigo R, Casanellas M. 2012. SPIn: model selection for phylogenetic mixtures via linear invariants. Mol. Biol. Evol. 29:929-937.

Kelly C. 1994. A test of the Markovian model of DNA evolution. Biometrics 50:653-664.

Kimura M. 1980. A Simple Method for Estimating Evolutionary Rates of Base Substitutions through Comparative Studies of Nucleotide-Sequences. J. Mol. Evol. 16:111-120.

Knight R, Maxwell P, Birmingham A, Carnes J, Caporaso JG, Easton BC, Eaton M, Hamady M, Lindsay H, Liu Z. 2007. PyCogent: a toolkit for making sense from sequence. Genome biology 8:R171.

Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K. 2012. Statistics and truth in phylogenomics. Mol. Biol. Evol. 29:457-472.

Kumar S, Gadagkar SR. 2001. Disparity index: a simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. Genetics 158:1321-1327.

Lake JA. 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc Natl Acad Sci U S A 91:1455-1459.

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.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25:1307-1320.

Lemmon AR, Milinkovitch MC. 2002. The metapopulation genetic algorithm: An efficient solution for the problem of large phylogeny estimation. Proc Natl Acad Sci U S A 99:10516-10521.

Lemmon EM, Lemmon AR. 2013. High-Throughput Genomic Data in Systematics and Phylogenetics. Annual Review of Ecology, Evolution, and Systematics, Vol 44 44:99-+.

Lockhart PJ, Steel MA, Hendy MD, Penny D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11:605-612.

Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. Mol. Biol. Evol. 19:1-7.

Martijn J, Vosseberg J, Guy L, Offre P, Ettema TJ. 2018. Deep mitochondrial origin outside the sampled alphaproteobacteria. Nature.

Maurin O, Anest A, Bellot S, Biffin E, Brewer G, Charles-Dominique T, Cowan RS, Dodsworth S, Epitawalage N, Gallego B, et al. 2021. A nuclear phylogenomic study of the angiosperm order Myrtales, exploring the potential and limitations of the universal Angiosperms353 probe set. Am. J. Bot. 108:1087-1111.

Minh BQ, Dang CC, Vinh LS, Lanfear R. 2021. QMaker: Fast and Accurate Method to Estimate Empirical Models of Protein Evolution. Syst. Biol. 70:1046-1060.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. 37:1530-1534.

Murray S, Jørgensen MF, Ho SY, Patterson DJ, Jermiin LS. 2005. Improving the analysis of dinoflagellate phylogeny based on rDNA. Protist 156:269-286.

Nabholz B, Kunstner A, Wang R, Jarvis ED, Ellegren H. 2011. Dynamic evolution of base composition: causes and consequences in avian phylogenomics. Mol. Biol. Evol. 28:2197-2210.

Naser-Khdour S, Minh BQ, Lanfear R. 2021. Assessing Confidence in Root Placement on Phylogenies: An Empirical Study Using Non-Reversible Models for Mammals. Syst. Biol.

Naser-Khdour S, Minh BQ, Zhang W, Stone EA, Lanfear R. 2019. The Prevalence and Impact of Model Violations in Phylogenetic Analysis. Genome Biol Evol 11:3341–3352.

Nesnidal MP, Helmkampf M, Bruchhaus I, Hausdorf B. 2010. Compositional Heterogeneity and Phylogenomic Inference of Metazoan Relationships. Mol. Biol. Evol. 27:2095-2104.

Paton T, Haddrath O, Baker AJ. 2002. Complete mitochondrial DNA genome sequences show that modern birds are not descended from transitional shorebirds. Proceedings of the Royal Society of London B: Biological Sciences 269:839-846.

Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005. Phylogenomics. Annu Rev Ecol Evol S 36:541-562.

Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21:1455-1458.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. PloS one 5:e9490.

Ran JH, Shen TT, Wang MM, Wang XQ. 2018. Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between Gnetales and angiosperms. Proc Biol Sci 285:20181012.

Reeves JH. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. J. Mol. Evol. 35:17-31.

Roberts D, Yang Z. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol. Biol. Evol. 12:451-458.

Rokas A, Chatzimanolis S. 2008. From gene-scale to genome-scale phylogenetics: the data flood in, but the challenges remain. In. Phylogenomics: Springer. p. 1-12.

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

Rzhetsky A, Nei M. 1995. Tests of applicability of several substitution models for DNA sequence data. Mol. Biol. Evol. 12:131-151.

Schwarz G. 1978. Estimating the dimension of a model. The annals of statistics:461-464.

Sheffield NC, Song H, Cameron SL, Whiting MF. 2009. Nonstationary Evolution and Compositional Heterogeneity in Beetle Mitochondrial Phylogenomics. Syst. Biol. 58:381-394.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492-508.

Squartini F, Arndt PF. 2008. Quantifying the stationarity and time reversibility of the nucleotide substitution process. Mol. Biol. Evol. 25:2525-2535.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312-1313.

Steel MA, Lockhart PJ, Penny D. 1993. Confidence in evolutionary trees from biological sequence data. Nature 364:440-442.

Stiller J, Tilic E, Rousset V, Pleijel F, Rouse GW. 2020. Spaghetti to a Tree: A Robust Phylogeny for Terebelliformia (Annelida) Based on Transcriptomes, Molecular and Morphological Data. Biology (Basel) 9:73.

Stuart A. 1955. A Test for Homogeneity of the Marginal Distributions in a Two-Way Classification. Biometrika 42:412-416.

Sullivan J, Joyce P. 2005. Model selection in phylogenetics. Annual Review of Ecology Evolution and Systematics 36:445-466.

Sumner JG, Fernandez-Sanchez J, Jarvis PD. 2012. Lie Markov models. J. Theor. Biol. 298:16-31.

Susko E, Roger AJ. 2020. On the use of information criteria for model selection in phylogenetics. Mol. Biol. Evol. 37:549-562.

Swofford DL. 2001. Paup*: Phylogenetic analysis using parsimony (and other methods) 4.0. B5.

Tamura K, Kumar S. 2002. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. Mol. Biol. Evol. 19:1727-1736.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10:512-526.

Tarrío R, Rodríguez-Trelles F, Ayala FJ. 2001. Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae. Mol. Biol. Evol. 18:1464-1473.

Tavaré S. 1986. Some probabilistic and statistical probles in the analysis of DNA sequences. Lectures on Mathematics in the Life Sciences 17.

Weiss G, von Haeseler A. 2003. Testing Substitution Models Within a Phylogenetic Tree. Mol. Biol. Evol. 20:572-578.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18:691-699.

Woodhams MD, Fernandez-Sanchez J, Sumner JG. 2015. A New Hierarchy of Phylogenetic Models Consistent with Heterogeneous Substitution Rates. Syst. Biol. 64:638-650.

Wu S, Edwards S, Liu L. 2018. Genome-scale DNA sequence data and the evolutionary history of placental mammals. Data Brief 18:1972-1975.

Yang Z. 1994. Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39:105-111.

Yang ZH, Roberts D. 1995. On the Use of Nucleic-Acid Sequences to Infer Early Branchings in the Tree of Life. Mol. Biol. Evol. 12:451-458.

Zou L, Susko E, Field C, Roger AJ. 2012. Fitting nonstationary general-time-reversible models to obtain edge-lengths and frequencies for the Barry–Hartigan model. Syst. Biol. 61:927-940.

CHAPTER 1

THE PREVALENCE AND IMPACT OF MODEL VIOLATIONS IN PHYLOGENETIC ANALYSIS

Suha Naser-Khdour*[1], Bui Quang Minh[1,2], Wenqi Zhang[1], Eric A. Stone[1], Robert Lanfear[1]

(1)Department of Ecology and Evolution, Research School of Biology, Australian National University, Canberra, Australian Capital Territory, Australia

(2) Research School of Computer Science, Australian National University, Canberra, Australian Capital Territory, Australia

*Author for Correspondence: E-mail: suha.naser@anu.edu.au

Contributions:

Suha Naser-Khdour (SNK) wrote the python script, performed the analysis, analysed and interpreted the results, drafted the manuscript, and submitted the article for publication. Bui Quang Minh (MB) contributed to the research design, conceptual development, editorial comments and implemented the tests in IQ-TREE. Wenqi Zhang (WZ) contributed to the research design and conceptual development. Eric A. Stone (ES) contributed to the research design and conceptual development. Robert Lanfear (RL) contributed to the research design, conceptual development and editorial comments.

Abstract

In phylogenetic inference, we commonly use models of substitution which assume that sequence evolution is stationary, reversible and homogeneous (SRH). Although the use of such models is often criticized, the extent of SRH violations and their effects on phylogenetic inference of tree topologies and edge lengths are not well understood. Here, we introduce and apply the maximal matched-pairs tests of homogeneity to assess the scale and impact of SRH model violations on 3,572 partitions from 35 published phylogenetic datasets. We show that roughly one-quarter of all the partitions we analysed (23.5%) reject the SRH assumptions and that for 25% of datasets, the topologies of trees inferred from all partitions differ significantly from those inferred using the subset of partitions that do not reject the SRH assumptions. This proportion of significantly different topologies is actually even greater when evaluating trees inferred using the subset of partitions that rejects the SRH assumptions, as compared to trees inferred from all partitions. These results suggest that the extent and effects of model violation in phylogenetics may be substantial. They highlight the importance of testing for model violations and possibly excluding partitions that violate models prior to tree reconstruction. Our results also suggest that further effort in developing models that do not require SRH assumptions could lead to large improvements in the accuracy of phylogenomic inference. The scripts necessary to perform the analysis are available in https://github.com/roblanf/SRHtests, and the new tests we describe are available as a new option in IQ-TREE (http://www.iqtree.org).

Keywords: model violations, phylogenetic inference, test of symmetry, systematic bias

Introduction

Phylogenetics is an essential tool for inferring evolutionary relationships between individuals, species, genes, and genomes. Moreover, phylogenetic trees form the basis of a huge range of other inferences in evolutionary biology, from gene function prediction to drug development and forensics (Eisen 1998; Farrell, et al. 2000; Mäser, et al. 2001; Gardner, et al. 2002; Yao, et al. 2003; Grenfell, et al. 2004; Yao, et al. 2004; Salipante and Horwitz 2006; Gray, et al. 2009; Brady and Salzberg 2011; Dunn, et al. 2011).

Most phylogenetic studies use models of sequence evolution which assume that the evolutionary process follows stationary, reversible and homogeneous (SRH) conditions. Stationarity implies that the marginal frequencies of the nucleotides or amino acids are constant over time, reversibility implies that the evolutionary process is stationary and undirected (substitution rates between nucleotides or amino acids are equal in both directions), and homogeneity implies that the instantaneous substitution rates are constant along the tree or over an edge (Felsenstein 2004; Yang and Rannala 2012; Jermiin, et al. 2017). However, these simplifying assumptions are often violated by real data (Foster and Hickey 1999; Tarrío, et al. 2001; Paton, et al. 2002; Goremykin and Hellwig 2005; Murray, et al. 2005; Bourlat, et al. 2006; Hyman, et al. 2007; Sheffield, et al. 2009; Nesnidal, et al. 2010; Nabholz, et al. 2011; Martijn, et al. 2018). Such model violation may lead to systematic error that, unlike stochastic error, cannot be remedied simply by increasing the size of a dataset (Felsenstein 2004; Ho and Jermiin 2004; Jermiin, et al. 2004; Philippe, et al. 2005; Sullivan and Joyce 2005; Kumar, et al. 2012; Brown and Thomson 2017; Duchene, et al. 2017). As phylogenetic datasets are steadily growing in terms of taxonomic and site sampling, it is vital that we develop and employ methods to measure and understand the extent to which systematic error affects phylogenetic inference (systematic bias), and explore ways of mitigating this systematic bias in empirical studies.

One approach to accommodate data that have evolved under non-SRH conditions is to employ models that relax the SRH assumptions. A number of non-SRH models have been implemented in a variety of software packages (Foster 2004; Lartillot and Philippe 2004; Blanquart and Lartillot 2006; Boussau and Gouy 2006; Jayaswal, et al. 2007; Knight, et al. 2007; Dutheil and Boussau 2008; Jayaswal, et al. 2011; Sumner, et al. 2012; Zou, et al. 2012; Groussin, et al. 2013; Jayaswal, et al. 2014; Nguyen, et al. 2015; Woodhams, et al. 2015). However, such models remain infrequently used as searching for optimal phylogenetic trees under these models is computationally demanding (Betancur-R, et al. 2013) and the implementations are often not easy to use. As a result, the vast majority of empirical phylogenetic inferences rely on models that assume sequences have evolved under SRH conditions, such as the general time-reversible (GTR) family of models implemented in many of the most widely-used phylogenetics software packages (Swofford 2001; Drummond and Rambaut 2007; Guindon, et al. 2010; Ronquist, et al. 2012; Bazinet, et al. 2014; Bouckaert, et al. 2014; Stamatakis 2014; Nguyen, et al. 2015; Höhna, et al. 2016).

Another approach to account for data that may have evolved under non-SRH conditions is to test for model violations prior to tree reconstruction. Here, one first screens datasets or parts of datasets, and reconstructs trees exclusively from data that do not reject SRH conditions. A number of methods have been proposed to test for violation of SRH conditions in aligned sequences prior to estimating trees (Bowker 1948; Stuart 1955; Rzhetsky and Nei 1995; Kumar and Gadagkar 2001; Weiss and von Haeseler 2003; Ababneh, et al. 2006; Ho, et al. 2006), and there are also *a posteriori* tests for absolute model adequacy which are employed after trees have been estimated (Goldman 1993; Bollback 2002; Brown and ElDabaje 2009; Brown 2014; Duchene, et al. 2017; Brown and Thomson 2018).

Allowing the data to reject the model when the assumptions of the model are violated is an important approach to reducing systematic bias in phylogenetic inference (Philippe, et al. 2005; Brown 2014). Knowing in advance which sequences and loci are inconsistent with the SRH assumptions will allow us to choose more complex models or to omit some of these sequences and loci from downstream analyses (Kumar and Gadagkar 2001). The need for methods that assess the evolutionary process prior to phylogenetic inference becomes more important as the number of sequences and sites per dataset increases because systematic bias has an increasing effect on inferences from larger phylogenetic datasets (Ho and Jermiin 2004; Jermiin, et al. 2004; Phillips, et al. 2004; Delsuc, et al. 2005).

In this paper, we evaluate the extent and effect of model violation due to non-SRH evolution using 35 empirical datasets with a total of 3,572 partitions. We determine if the SRH assumptions are violated by extending and applying the matched-pairs tests of homogeneity (Jermiin, et al. 2017) to each partition. We then compare the phylogenetic trees for each dataset estimated from all of the partitions, the partitions that reject the SRH assumptions, and the partitions that do not reject the SRH assumptions, in order to evaluate the effect of violating SRH conditions on phylogenetic inference. Our results suggest that violating SRH assumptions can have substantial impacts on phylogenetic inference.

Materials and Methods

Empirical datasets

In order to assess the impact of model violation in phylogenetics, we first gathered a representative sample of 35 partitioned empirical datasets that had been used for phylogenetic analysis in recent studies (Table 1). Within the constraints of selecting data that were publicly available and suitably annotated, i.e. such that all loci and all codon positions within protein-coding loci could be identified, we selected the datasets to provide as representative a sample as possible of the data types, taxa, and genomic regions most commonly used to infer bifurcating phylogenetic trees from concatenated alignments. These datasets include nucleotide sequences from nuclear, mitochondrial, plastid and virus genomes, and include protein-coding DNA, introns, intergenic spacers, tRNA, rRNA and ultra-conserved elements. The number of taxa and sites in these datasets range from 27 to 355 and from 699 to 1,079,052 respectively. The clades represented in these datasets include animals, plants and viruses. We partitioned all datasets to the maximum possible extent based on the biological properties of the data, i.e. we divided every locus and every codon position within each protein-coding locus into a separate partition. All partitioning information is available at the github repository https://github.com/roblanf/SRHtests/tree/master/datasets, and the full details of every dataset are provided in Table 1 and in extended Table 5.

Table 1| Number of taxa, number of sites, clade and study reference for each dataset that has been used in this study

Dataset	Study Reference	Dataset Reference	Clade	Taxa	Sites
Anderson_2013	(Anderson, et al. 2014)	(Anderson, et al. 2013)	loliginids	145	3037
Bergsten_2013	(Bergsten, et al. 2013a)	(Bergsten, et al. 2013b)	Dytiscidae	38	2111
Broughton_2013	(Broughton, et al. 2013b)	(Broughton, et al. 2013a)	Osteichthyes	61	19997
Brown_2012	(Brown, et al. 2012b)	(Brown, et al. 2012a)	Ptychozoon	41	1665
Cannon_2016a	(Cannon, et al. 2016a)	(Cannon, et al. 2016b)	Metazoa	78	89792
Cognato_2001	(Cognato and Vogler 2001b)	(Cognato and Vogler 2001a)	Coleoptera: Scolytinae	44	1897

Day_2013	(Day, Peart, Brown, Friel, et al. 2013)	(Day, Peart, Brown, Bills, et al. 2013)	Synodontis	152	3586
Devitt_2013	(Devitt, Devitt, et al. 2013)	(Devitt, Cameron Devitt, et al. 2013)	Ensatina eschscholtzii klauberi	69	823
Dornburg_2012	(Dornburg, et al. 2012b)	(Dornburg, et al. 2012a)	Teleostei: Beryciformes: Holocentridae	44	5919
Faircloth_2013	(Faircloth, et al. 2013b)	(Faircloth, et al. 2013a)	Actinopterygii	27	149366
Fong_2012	(Fong, et al. 2012b)	(Fong, et al. 2012a)	Vertebrata	110	25919
Horn_2014	(Horn, et al. 2014b)	(Horn, et al. 2014a)	Euphorbia	197	11587
Kawahara_2013	(Kawahara and Rubinoff 2013a)	(Kawahara and Rubinoff 2013b)	Hyposmocoma	70	2238
Lartillot_2012	(Lartillot and Delsuc 2012b)	(Lartillot and Delsuc 2012a)	Eutheria	78	15117
McCormack_2013	(McCormack, et al. 2013b)	(McCormack, et al. 2013a)	Neoaves	33	1079052
Moyle_2016	(Moyle, et al. 2016b)	(Moyle, et al. 2016a)	Oscines	106	375172
Murray_2013	(Murray, et al. 2013a)	(Murray, et al. 2013b)	Eucharitidae	237	3111
Oaks_2011	(Oaks 2011b)	(Oaks 2011a)	Crocodylia	79	7282
Rightmyer_2013	(Rightmyer, et al. 2013b)	(Rightmyer, et al. 2013a)	Hymenoptera: Megachilidae	94	3692
Sauquet_2011	(Sauquet, et al. 2012)	(Sauquet, et al. 2011)	Nothofagus	51	5444
Seago_2011	(Seago, et al. 2011b)	(Seago, et al. 2011a)	Coccinellidae	97	2253
Sharanowski_2011	(Sharanowski, et al. 2011b)	(Sharanowski, et al. 2011a)	Braconidae	139	3982
Siler_2013	(Siler, Oliveros, et al. 2013)	(Siler, Brown, et al. 2013)	Lycodon	61	2697
Tolley_2013	(Tolley, et al. 2013b)	(Tolley, et al. 2013a)	Chamaeleonidae	203	5054
Unmack_2013	(Unmack, et al. 2013b)	(Unmack, et al. 2013a)	Melanotaeniidae	139	6827
Wainwright_2012	Wainwright, Smith, Price, Tang, Sparks, Ferry, Kuhn, Eytan, et al. (2012)	(Wainwright, Smith, Price, Tang, Sparks, Ferry, Kuhn and Near 2012)	Acanthomorpha	188	8439
Wood_2012	(Wood, et al. 2013)	(Wood, et al. 2012)	Archaeidae	37	5185
Worobey_2014a	(Worobey, et al. 2014b)	(Worobey, et al. 2014a)	Influenzavirus A	146	3432
Worobey_2014b	(Worobey, et al. 2014b)	(Worobey, et al. 2014a)	Influenzavirus A	327	759
Worobey_2014c	(Worobey, et al. 2014b)	(Worobey, et al. 2014a)	Influenzavirus A	92	1416
Worobey_2014d	(Worobey, et al. 2014b)	(Worobey, et al. 2014a)	Influenzavirus A	355	1497
Worobey_2014e	(Worobey, et al. 2014b)	(Worobey, et al. 2014a)	Influenzavirus A	340	699
Worobey_2014f	(Worobey, et al. 2014b)	(Worobey, et al. 2014a)	Influenzavirus A	332	2151
Worobey_2014g	(Worobey, et al. 2014b)	(Worobey, et al. 2014a)	Influenzavirus A	326	2274
Worobey_2014h	(Worobey, et al. 2014b)	(Worobey, et al. 2014a)	Influenzavirus A	351	2280

Workflow summary

Figure 1 outlines the workflow. For each partition in each dataset, we used a new approach based on the three matched-pairs tests of homogeneity to ask whether the evolution of the aligned sequences in the partition rejects the SRH assumptions. The three matched-pairs tests of homogeneity, described in more detail below, test three slightly different assumptions about the historical process that generated each aligned pair of sequences in a given partition. A significant result from any test suggests that the nature of the evolutionary process required to explain the aligned sequences violates at least one of the three SRH conditions (Jermiin, et al. 2017). For each test, we classify each partition as *pass* if the result of the test is non-significant or *fail* if the result of the test is significant. We then denote the original dataset as D_{all}, while the concatenation of *pass* partitions is denoted D_{pass} and the concatenation of *fail* partitions as D_{fail} (fig. 1).

To investigate the impact of model violation on phylogenetic inference, we infer and compare three phylogenetic trees, T_{all}, T_{pass} and T_{fail}, estimated from D_{all}, D_{pass} and D_{fail}, respectively.

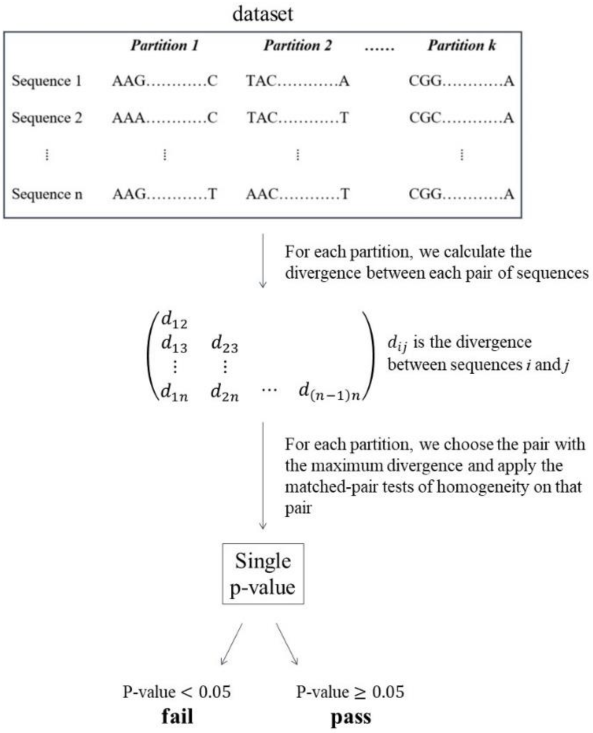

dataset
Partition 1
Partition 2
……
Partition k
Sequence 1
AAG...........C
TAC...........A
CGG...........A
Sequence 2
AAA...........C
TAC...........T
CGC...........A
Sequence n
AAG...........T
AAC...........T
CGG...........A
For each partition, we calculate the divergence between each pair of sequences
d_{12} d_{13} d_{23} d_{1n} d_{2n} ⋯ $d_{(n-1)n}$
d_{ij} is the divergence between sequences i and j
For each partition, we choose the pair with the maximum divergence and apply the matched-pair tests of homogeneity on that pair
Single p-value
P-value < 0.05
fail
P-value ≥ 0.05
pass

Fig. 1| Flow chart of methodology. For each partition in the alignment, we choose the pair of sequences with the maximum divergence and apply the matched-pairs tests of homogeneity on that pair.

Matched-pairs tests of homogeneity

The three matched-pairs tests of homogeneity that are applied to pairs of sequences are: the MPTS (matched-pairs test of symmetry), MPTMS (matched-pairs test of marginal symmetry), and MPTIS (matched-pairs test of internal symmetry). The statistics are computed on an m-by-m (m is 4 for nucleotides and 20 for amino acids) divergence matrix D with elements d_{ij}, where d_{ij} is the number of alignment sites having nucleotide (or amino acid) i in the first sequence and nucleotide (or amino acid) j in the second sequence.

The MPTS tests the symmetry of D by computing the Bowker's (Bowker 1948) test statistic as the chi-square distance between D and its transpose:

$$S_B^2 = \sum_{1 \le i < j \le m} \frac{(d_{ij} - d_{ji})^2}{(d_{ij} + d_{ji})}$$

Where $d_{ij} + d_{ji} > 0$. A p-value is then obtained by a chi-square test with f degrees of freedom, where f is the number of (i, j) pairs for which $d_{ij} + d_{ji} > 0$. A small p-value (e.g. <0.05) indicates that the assumption of symmetry is rejected at that significance level, suggesting that evolution is non-stationary, non-homogeneous or both (Jermiin, et al. 2017).

The MPTMS tests the equality of nucleotide or amino acid composition between two sequences. To do so, MPTMS computes the Stuart's test statistic $S_S^2 = u^T V^{-1} u$ using the difference between nucleotide or amino acid frequencies of two sequences, u, and its variance-covariance matrix, V. In detail, u is given by $u^T = (d_{1\bullet} - d_{\bullet 1}, d_{2\bullet} - d_{\bullet 2}, \dots, d_{k\bullet} - d_{\bullet k})$ where $d_{i\bullet}$ is the sum of d_{ij} over j, $d_{\bullet j}$ is the sum of d_{ij} over i, and, k = m -1. V, the estimated variance-covariance matrix of u under the assumption of marginal symmetry, is defined elementwise by

$$v_{ij} = \begin{cases} d_{i\bullet} + d_{\bullet i} - 2d_{ii}, & \mathrm{i = j} \\ -(d_{ij} + d_{ji}), & \mathrm{i \neq j} \end{cases}$$

A p-value is obtained by a chi-square test with $m - 1$ degrees of freedom. A small p-value (<0.05) indicates that the stationarity assumption is rejected. Note that when V is not invertible, the Stuart's statistic S_S^2 is ill-defined and the MPTMS is not applicable.

The MPTIS uses the test statistic as the difference between Bowker's and Stuart's statistic: $S_I^2 = S_B^2 - S_S^2$. S_I^2 is chi-square distributed with $f - m + 1$ degrees of freedom. A small p-value (<0.05) indicates that the homogeneity assumption is rejected.

The MPTS, MPTMS and MPTIS test different aspects of the symmetry with which differences accumulate between pairs of sequences due to the substitution process. The MPTS is a comprehensive and sufficient test to determine whether the data complies with the SRH assumptions (Jermiin, et al. 2017), but it cannot provide any information about the source of this violation. Some information on the underlying source of model violation may be obtained by performing the other two tests of symmetry: the MPTMS and the MPTIS. If the violation of the SRH assumptions stems from differences in base composition between the sequences, this should affect the marginal symmetry of the sequence pair, which can in principle be detected by the MPTMS. If the violation of the SRH assumptions stems from changes in the relative substitution rates over time, this should affect the internal symmetry of the sequence pair, which can in principle be detected by the MPTIS. However, even after performing all three tests, it is difficult to ascertain which of the three SRH assumptions is violated during the evolutionary process because the relationships between the SRH conditions and the three matched-pair tests is neither bijective nor injective, i.e. there is not a one-to-one correspondence between the three tests and violation of the three SRH conditions (Jermiin, et al. 2017).

The three matched-pairs tests of homogeneity are appropriate to test for SRH assumptions as they consider the alignment on a site-by-site basis. The basic intuition that underlies these tests is that two sequences diverging under SRH conditions should accumulate differences symmetrically (e.g. both sequences are equally likely to accumulate at a C to T change at a site in which both originally shared a C). This symmetry of accumulation is reflected by symmetries in the resulting difference matrix, violations of which can be assessed statistically. However, these tests were designed to ask whether any single pair of sequences rejects the SRH conditions (Jermiin, et al. 2017). To ask whether a given partition rejects SRH conditions, we developed an approach to extend the matched-pairs tests of homogeneity to accommodate datasets with more than two sequences.

Maximum Symmetry Test

In order to determine whether a given multiple sequence alignment rejects SRH conditions, we consider only the pair of taxa with the maximum divergence. In order to find the maximum divergent pair, we sum the off-diagonal elements of the divergence matrix and divide by the sum of all elements. We then randomly choose one pair from all the pairs with the maximum divergence score (if there are more than one pair). By using the most divergent sequence pair, we maximise our power to detect model violations without a priori knowledge of the underlying tree topology and the dependencies that it induces in the data. For the maximum divergent pair, we then apply the matched-pair tests of homogeneity and calculate their chi-squared p-values. If the obtained p-value is less than 0.05, then we consider that the null hypothesis of SRH evolution is rejected for the corresponding partition and we add it to the D_{fail} dataset. Otherwise, we add it to the D_{pass} dataset. We denote our applications of the MPTS, MPTMS, and MPTIS based on the $d_{max}\ Pair$ as MaxSymTest, MaxSymTest_{mar}, and MaxSymTest_{int}, respectively.

Phylogenetic inference

We used IQ-TREE (Nguyen, et al. 2015) to infer up to seven phylogenetic trees for every dataset: T_{all} (all partitions from the original dataset; D_{all}); and T_{pass} and T_{fail} based on the D_{pass} and D_{fail} datasets from each of the three tests (MaxSymTest, $MaxSymTest_{mar}$, $MaxSymTest_{int}$), provided that there was at least one partition in each category. We ran IQ-TREE using the default settings with the best-fit fully-partitioned model (Chernomor, et al. 2016), which allows each partition to have its own evolutionary model and edge-linked rate determined by ModelFinder (Kalyaanamoorthy, et al. 2017) followed 1000 ultrafast bootstrap replicates (Hoang, et al. 2018).

Distance between trees

For each of the three tests (MPTS, MPTMS, MPTIS) we calculated the Normalised Path-Difference (NPD) and quartet distance (QD) (Steel and Penny 1993; Sand, et al. 2014) between all three possible pairs of trees (T_{all} vs. T_{pass}; T_{all} vs. T_{fail}; and T_{pass} vs. T_{fail}), as long as D_{pass} and D_{fail} were non-empty and so T_{pass} and T_{fail} had been estimated. The path-difference metric (PD) is defined as the Euclidean distance between pairs of taxa (Steel and Penny 1993; Mir and Russello 2010). In this study, because we are interested only in differences between topologies, we use the variant of the PD metric that ignores branch lengths. In order to compare path distances between trees with different number of taxa, we normalised PD (to obtain NPD) by the mean of a null distribution of PDs generated from 10K random pairs of trees with the same number of taxa (Bogdanowicz, et al. 2012). Thus, an NPD of zero indicates an identical pair of trees, an NPD of 1 indicates that a pair of trees is as similar as a pair of randomly-selected trees with the same number of taxa; and an NPD greater than 1 indicates a pair of trees that are less similar than a randomly-selected pair of trees with the same number of taxa. Since path differences are always non-negative, the NPD is also guaranteed to be non-negative.

The QD metric is defined as the fraction of quartets (subsets of four taxa) that induce different subtrees between the two trees being compared. QD ranges between 0 and 1, where 0 means that two trees are identical and 1 means that they do not share any quartet subtrees. Compared with PD, QD has the advantage that its distribution is less sensitive to the underlying distribution of tree topologies (Steel and Penny 1993).

Tree topology tests

The NPD and the QD give us measures of the differences between pairs of trees, but they do not tell us whether the differences are phylogenetically significant in the three datasets (D_{pass}, D_{all}, and D_{fail}) derived from a given test. For example, trees that differ due to stochastic error associated with small datasets may be very different, but such differences may not be statistically significant. To assess the significance of the differences between T_{pass}, T_{all} and T_{fail}, we used the weighted Shimodaira-Hasegawa (wSH) test (Shimodaira and Hasegawa 1999; Shimodaira 2002) implemented in IQ-TREE with 1000 RELL replicates (Kishino, et al. 1990). Given the alignment (D_{pass}), the wSH test computes a p-value for each tree, where a small p-value (<0.05) implies that the corresponding tree has a significantly worse likelihood than the best tree in the set of T_{pass}, T_{all} and T_{fail}. We use D_{pass} for these tests because it is, by definition, the only dataset that does not reject the underlying assumptions of the SH test. As such, we only compute sWH p-values when D_{pass} is non-empty. Thus, we performed a wSH test for each of the three MaxSymTest variants: each of which asks whether T_{all} and/or T_{fail} can be rejected in favour of T_{pass}.

Correlation between number of substitutions and model violation

We hypothesised that partitions with more substitutions may be more likely to violate the SRH assumptions since substitutions form the raw data for the matched-pairs tests of homogeneity. To assess this, we fitted a linear mixed-effects model for each of the three tests using the glmer function from the lme4 package in R (Bates, et al. 2015). In this model, we treat each partition as a datapoint, the number of substitutions measured for that partition as a fixed effect, and the dataset from which that partition was taken as a random effect. This allows us to estimate the extent to which the number of substitutions in a partition associate with whether a partition fails a given test of symmetry, after accounting for differences between the datasets. To calculate the R-squared value we use the r.squaredGLMM function from the MuMIn package in R (Barton 2009; Nakagawa and Schielzeth 2013).

Software implementation

We implemented a new option --symtest in IQ-TREE to perform the three MaxSymTest matched pairs tests of symmetry. In addition, the option --symtest-remove-bad allows users to remove from the final analysis partitions that fail the MaxSymTest. One can change the removal criterion to MaxSymTest$_{mar}$ or MaxSymTest$_{int}$ via the --symtest-type MAR|INT option. In addition, the cut-off p-value can be changed using the --symtest-pval NUM option, where the default value is 0.05.

Reproducibility

The GitHub repository https://github.com/roblanf/SRHtests contains the raw data and Python and R scripts necessary to perform all analyses reported in this study.

Results

Violation of SRH conditions is common across 35 empirical datasets

Across all 3,572 partitions analysed, 573 (16.0%) failed the MaxSymTest, 728 (20.4%) failed the MaxSymTest$_{mar}$, and 312 (2.8%) failed the MaxSymTest$_{int}$. In total, 840 (23.5%) of the partitions failed at least one test.

The proportion of partitions failing each test varied substantially among datasets (fig. 2), but on average 21.8% of the partitions in each dataset failed the MaxSymTest, 27.5% failed the $\text{MaxSymTest}_{\text{mar}}$, and 5.1% failed the $\text{MaxSymTest}_{\text{int}}$.

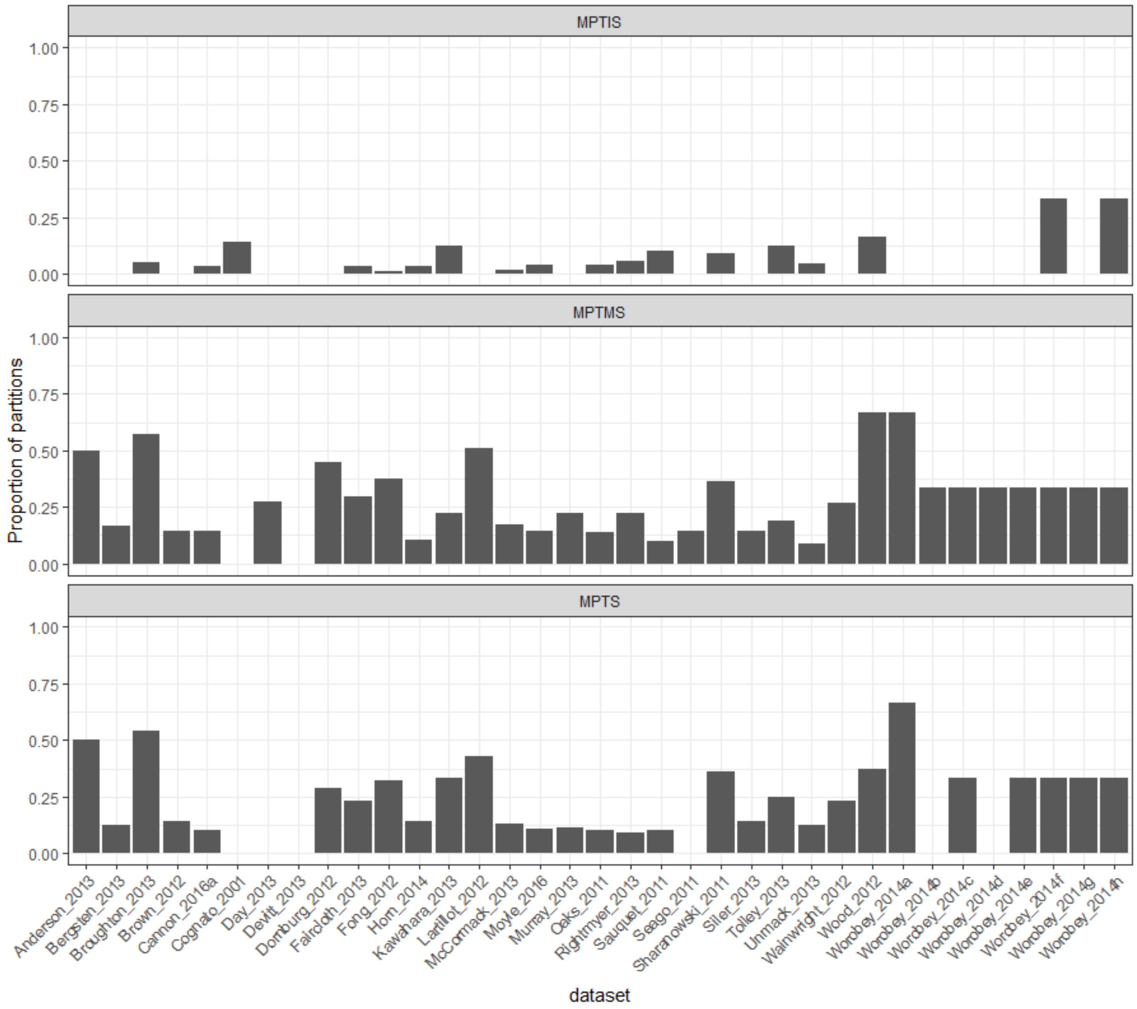


Fig. 2| The proportion of partitions that reject the null hypothesis of the MaxSymTest, $\text{MaxSymTest}_{\text{mar}}$ and $\text{MaxSymTest}_{\text{int}}$ (p-value < 0.05) in each dataset.

The fraction of failing partitions also varied with the genome type (e.g. mitochondrial, chloroplast, or nuclear) and context (e.g. protein-coding, UCE, tRNA) from which the partition was sequenced (Table 2) although we note that a substantial proportion of the partitions from almost every category failed at least one of the tests (Table 2).

There were no clear differences in the substitution models that were selected for the partitions that pass or fail the tests (see Extended Tables 1-3). However, we note that the two most frequently selected substitution models (for 35% of the partitions) were relatively simple: K80 (Kimura 1980) and HKY (Hasegawa, et al. 1985).

Table 2| The proportion of partitions that failed at least one of the three tests - MaxSymTest, MaxSymTest$_{mar}$, MaxSymTest$_{int}$

Type / genome	nuclear	mitochondrial	plastid	virus
1st codon positions	20.2%	27.6%	33.3%	25.0%
2nd codon positions	21.0%	7.4%	0.0%	25.0%
3rd codon positions	76.6%	44.8%	0.0%	75.0%
Other (e.g. intron)	27.8%	100.0%	0.0%	
rRNA	30.0%	25.0%		
UCE	22.5%			
tRNA		0.0%		

Model violation has a large influence on tree topologies

Using both MaxSymTest and MaxSymTest$_{mar}$, we compared each tree inferred from each dataset (T_{all}) to the corresponding trees estimated from the failed (T_{fail}) and passed (T_{pass}) partitions. Disturbingly, for each of the two tree distance metrics that we considered (NPD and QD), we find that the tree inferred from the original dataset tended to be more similar to the tree estimated from the failed partitions (Table 3, Extended Table 4). Furthermore, the mean NPD distance between T_{pass} and T_{fail} across all 35 datasets for the MaxSymTest was 0.69, i.e., the two trees are 69% as dissimilar as random pairs of trees. This suggests that violations of SRH assumptions drive large changes in tree topologies.

Table 3| The proportion of datasets that have the highest NPD metric (and QD metric) between the three comparisons (all-fail, all-pass, pass-fail) for MaxSymTest, MaxSymTest$_{mar}$, and MaxSymTest$_{int}$.

		T_{fail}	T_{pass}
MaxSymTest	T_{all}	14.3% (4.8%)	4.8% (4.8%)
	T_{pass}	80.9% (90.4%)	
MaxSym Test$_{mar}$	T_{all}	8.3% (0.0%)	8.3% (4.2%)
	T_{pass}	83.4% (95.8%)	
MaxSym Test$_{int}$	T_{all}	28.6% (28.6%)	0.0% (0.0%)
	T_{pass}	71.4% (71.4%)	

The results of the wSH tests (Table 4) confirm that the differences between trees that we observe tend to be statistically significant. For example, when using the MaxSymTest$_{mar}$, T_{pass} is a significantly better description of the D_{pass} data than T_{all} in ~37% of the datasets, and better than T_{fail} in ~89% of the datasets.

Table 4| The proportion of datasets that have a significant p-value in the weighted SH test when using D_{pass} as the input alignment for the test.

	T_{all}	T_{fail}
MaxSymTest	25%	79%
MaxSymTest$_{mar}$	37%	89%
MaxSymTest$_{int}$	4%	28%

The number of substitutions explains less than one-third of the variance in passing or failing the tests of symmetry

The number of substitutions in a partition explained 27.5% of the variation in whether or not a partition passed or failed the MaxSymTest (Extended Fig. 7). This proportion is very similar for MaxSymTest$_{mar}$ (24.4%) (Extended Fig. 8) but is dramatically lower for the MaxSymTest$_{int}$ (1.8%) (Extended Fig. 9). Thus, although the number of substitutions in a partition is a highly significant ($p<2e-16$) predictor of passing or failing any of the tests, that it explains only about a quarter of the variation suggests that other factors, such as underlying differences in the extent to which partitions violate the SRH assumptions, are driving the remaining ~75% of the variation.

Model violation due to non-SRH evolution affects the inferred relationship between even-toed and odd-toed ungulates in the tree of mammals

To examine the effects of model violation in more detail, we selected two datasets for more detailed consideration. Conflicting support for the placement of Xenacoelomorpha, the clade that contains Xenoturbella and Acoelomorpha, in the tree of life across different analyses has led to various hypotheses about the evolution of Bilateria (Cannon, et al. 2016a). In addition, the interordinal relationships in Laurasiatheria, especially the relationships between Fereuungulata (Perissodactyla, Cetartiodactyla, Carnivora, and Pholidota), in the tree of placental mammals is controversial (Cao, et al. 1998; Zhou, et al. 2012). It has been suggested that such inferences might be strongly affected by model violation and systematic error (Cao, et al. 1998; Delsuc, et al. 2005; Philippe, et al. 2011; Tsagkogeorga, et al. 2013). To assess whether data that pass or fail the MaxSymTest$_{mar}$ show different signals regarding the evolution of the Bilateria and the superorder Laurasiatheria, we examined in more detail the T_{all} T_{pass} , and T_{fail} trees from recent studies that explored the tree of placental mammals (Lartillot and Delsuc 2012b) and the tree of all animals (Cannon, et al. 2016a). The mammals' dataset comprises 78 mammalian taxa, including 73 placental mammals with 5 partitions representing the first, second, and third codon positions of the 17 genes (Lartillot and Delsuc 2012a). The tree reconstructed from all of the partitions (T_{all}) and the tree reconstructed from the partitions that pass the MaxSymTest (T_{pass}, 29 partitions) both show Perissodactyla (odd-toed ungulates) as a sister group to Cetartiodactyla (even-toed ungulates) (fig. 3a, Extended figs. 4-5). Even so, the bootstrap support for this branch is not high: 73% for T_{all} and 34% for T_{pass}. On the other hand, the tree reconstructed from the data that fail the MaxSymTest (T_{fail}, 22 partitions) shows Perissodactyla as the sister group to the clade that contains Carnivora + Pholidota with 49% bootstrap support (fig. 3b, Extended Fig. 6).

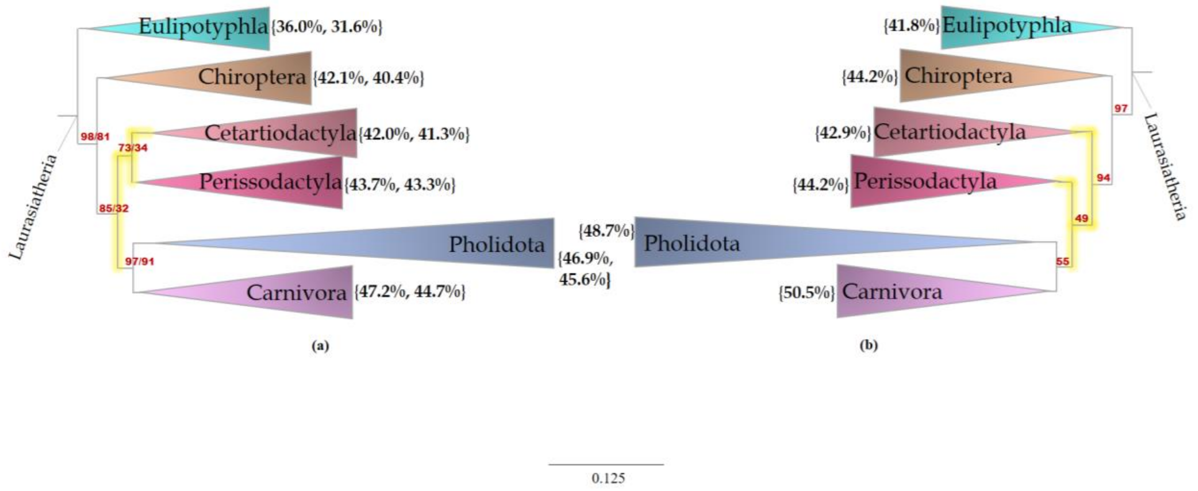


Fig. 3| Maximum likelihood trees of mammalian relationships based on analysis of Lartillot 2012 dataset. a) the tree inferred from all 51 partitions and from the 29 partitions that passed the MaxSymTest. b) the tree inferred from 22 partitions that failed the MaxSymTest. Red numbers at the internal branches indicate the bootstrap support values that are less than 100% under the best fitting model. Numbers in curly brackets show the GC content (in panel a, %GC and bootstrap support values are for T_{all} and T_{pass} respectively).

The animal dataset comprises 76 metazoan taxa, 2 choanoflagellate outgroups, 212 genes and 424 partitions representing first and second codon positions (Cannon, et al. 2016b). The tree reconstructed from all of the partitions (T_{all}) is identical to the trees reconstructed from the 381 partitions that pass the MaxSymTest (T_{pass}), the partitions that fail the MaxSymTest (T_{pass}, 43 partitions), and the tree shown in the original paper from both DNA and amino acid data

(Canon, et al. 2016a), which places Xenacoelomorpha as the sister group of Nephrozoa (Deuterostomia and Protostomia) with 100% bootstrap support (Extended figs. 1-3).

Discussion

In this paper, we show that model violation is prevalent and has a strong impact on tree reconstruction in many phylogenetic datasets. This impact varies substantially between different datasets and different types of partitions. The trees inferred from different groups of partitions from the same dataset often have topologies that are biologically and statistically significantly different.

Our results show great heterogeneity in the extent of model violation among different datasets and partitions. This is demonstrated by the varying proportion of partitions that failed the matched-pairs tests of homogeneity in each dataset and in each genomic context (codon position, rRNA, tRNA, UCE or other) and type of genome (nuclear, mitochondrial, plastid and virus). Model violations are most frequently observed in the third codon positions for viral, mitochondrial and nuclear genomes and intergenic spacers in plastid sequences. Yet, our results affirm that non-SRH evolution is far from constrained to these genomic regions. For example, in a dataset of placental mammals, of the 22 partitions that failed the MaxSymTest, only 11 are third codon positions. The tree inferred from the partitions that show a significant violation of the SRH conditions (T_{fail}) differs in its topology from the tree inferred from the partitions that do not show a significant violation of the SRH conditions (T_{pass}) with respect to the interordinal relationships in Laurasiatheria (fig. 3). The tree inferred from partitions that violate the SRH conditions (T_{fail}) is consistent with the results from the original paper in that it places Perissodactyla as a sister group to Carnivora + Pholidota (Lartillot and Delsuc 2012b). However, other studies using ML analysis show Perissodactyla to be a sister group to

Cetartiodactyla (Graur, et al. 1997; Murphy, et al. 2001; Tsagkogeorga; et al. 2013, Liu, et al. 2017), which is also the relationship we find in this study with the tree inferred from partitions that do not show a significant violation of the SRH assumptions.

Examining the results of the two other tests (MaxSymTest$_{mar}$ and MaxSymTest$_{int}$) we noticed that all the partitions that failed the MaxSymTest also failed the MaxSymTest$_{mar}$, suggesting that those partitions are violating the models mainly due to non-stationarity. Based on this observation, GC content may drive the differences between the trees inferred from all partitions and those inferred from partitions that failed neither MaxSymTest nor MaxSymTest$_{mar}$. Trees with partitions that violate the models tend to group together clades with similar GC content (e.g. as in Betancur-R, et al. 2013). However, it is hard to discern any clear evidence for this from examining the GC content of the clades (Figure 3). Yet, our results show that all the clades in the partitions that failed the MaxSymTest have on average a higher GC content (Figure 3).

The results of our study also provide some insight into the likely cause of model violation in the datasets we examined. Figure 2 shows that violation of marginal symmetry (assessed with MaxSymTest$_{mar}$) was much more common than violation of internal symmetry (assessed with MaxSymTest$_{int}$). This suggests that non-stationarity, which is associated with marginal symmetry, is likely a more common cause of systematic bias than non-homogeneity in the datasets that we examined (see also Jayaswal, et al. 2005; Ababneh, et al. 2006; Song, et al. 2010). Yet, the difference between the proportion of partitions that failed the MaxSymTest$_{mar}$ and the proportion of partitions that failed the MaxSymTest$_{int}$ could also be due to the higher power of the MaxSymTest$_{mar}$. Either way, this result hints that the development and application of non-stationary models (e.g. Yang 1994; Roberts and Yang 1995; Yap and Speed 2005) may be an important avenue towards reducing systematic bias in future analyses. Moreover, our results show a clear preference for simple substitution models with a single transition/transversion ratio over more complex models such as GTR. This suggests that developing non-stationary models with a single parameter for the transition/transversion ratio might be sufficient to reduce systematic bias in phylogenetic analysis.

One limitation of using the tests that we propose in this paper is that their power will be limited if there are few differences between the sequences being examined. Indeed, our analyses show that in our representative sample of more than 3500 partitions from published datasets, roughly ~25% of the variance in whether a partition passes or fails a given test can be attributed to the number of observed differences between the sequences. Nevertheless, this implies that the remaining ~75% of the variance in whether a partition passes or fails a test could be attributable to other processes, such as variation in the extent of model violation among partitions. This suggests that we should be cautiously optimistic: although a lack of power on small or slowly-evolving partitions may induce some false negatives (i.e. failures to identify partitions that have evolved under non-SRH conditions), the tests we propose still have significant power to identify partitions that show the evidence of model violation. It is possible that removing such partitions from phylogenetic analyses may improve the accuracy of results by reducing the overall burden of model violation on the inference of the tree topology. We hope that our implementation of these tests in the user-friendly software IQ-TREE will allow empirical phylogeneticists to continue to explore whether this is the case.

Acknowledgements

The authors would like to thank Lars Jermiin, David Bryant, Jeremy Brown and one anonymous referee for providing thoughtful comments on this manuscript.

References

Ababneh F, Jermiin LS, Ma C, Robinson J. 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics 22:1225-1231.

Anderson FE, Bergman A, Cheng SH, Pankey MS, Valinassab T. 2013. Data from: Lights out: the evolution of bacterial bioluminescence in Loliginidae. In: Dryad Data Repository.

Anderson FE, Bergman A, Cheng SH, Pankey MS, Valinassab T. 2014. Lights out: the evolution of bacterial bioluminescence in Loliginidae. Hydrobiologia 725:189-203.

Barton K. 2009. MuMIn: multi-model inference, R package version 0.12. 0. http://r-forge. r-project. org/projects/mumin/.

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting Linear Mixed-Effects Models Using lme4. 2015 67:48.

Bazinet AL, Zwickl DJ, Cummings MP. 2014. A gateway for phylogenetic analysis powered by grid computing featuring GARLI 2.0. Syst. Biol. 63:812-818.

Bergsten J, Nilsson AN, Ronquist F. 2013a. Bayesian tests of topology hypotheses with an example from diving beetles. Syst. Biol. 62:660-673.

Bergsten J, Nilsson AN, Ronquist F. 2013b. Data from: Bayesian tests of topology hypotheses with an example from diving beetles. In: Dryad Data Repository.

Betancur-r R, Li C, Munroe TA, Ballesteros JA, Ortí G. 2013. Addressing gene tree discordance and non-stationarity to resolve a multi-locus phylogeny of the flatfishes (Teleostei: Pleuronectiformes). Syst. Biol. 62:763-785.

Blanquart S, Lartillot N. 2006. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Mol. Biol. Evol. 23:2058-2071.

Bogdanowicz D, Giaro K, Wrobel B. 2012. TreeCmp: Comparison of Trees in Polynomial Time. Evol Bioinform 8:475-487.

Bollback JP. 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. 19:1171-1180.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comp. Biol. 10:e1003537.

Bourlat SJ, Juliusdottir T, Lowe CJ, Freeman R, Aronowicz J, Kirschner M, Lander ES, Thorndyke M, Nakano H, Kohn AB. 2006. Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. Nature 444:85.

Boussau B, Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. Syst. Biol. 55:756-768.

Bowker AH. 1948. A test for symmetry in contingency tables. J Am Stat Assoc 43:572-574.

Brady A, Salzberg S. 2011. PhymmBL expanded: confidence scores, custom databases, parallelization and more. Nat. Methods 8:367.

Broughton RE, Betancur RR, Li C, Arratia G, Orti G. 2013a. Data from: Multi-locus phylogenetic analysis reveals the pattern and tempo of bony fish evolution. In: Dryad Data Repository.

Broughton RE, Betancur RR, Li C, Arratia G, Orti G. 2013b. Multi-locus phylogenetic analysis reveals the pattern and tempo of bony fish evolution. PLoS Curr 5.

Brown JM. 2014. Detection of Implausible Phylogenetic Inferences Using Posterior Predictive Assessment of Model Fit. Syst. Biol. 63:334-348.

Brown JM, ElDabaje R. 2009. PuMA: Bayesian analysis of partitioned (and unpartitioned) model adequacy. Bioinformatics 25:537-538.

Brown JM, Thomson RC. 2017. Bayes Factors Unmask Highly Variable Information Content, Bias, and Extreme Influence in Phylogenomic Analyses. Syst. Biol. 66:517-530.

Brown JM, Thomson RC. 2018. Evaluating Model Performance in Evolutionary Biology. Annu Rev Ecol Evol S 49:null.

Brown RM, Siler CD, Das I, Min PY. 2012a. Data from: Testing the phylogenetic affinities of Southeast Asia’s rarest geckos: Flap-legged geckos (Luperosaurus), Flying geckos (Ptychozoon) and their relationship to the pan-Asian genus Gekko. In: Dryad Data Repository.

Brown RM, Siler CD, Das I, Min Y. 2012b. Testing the phylogenetic affinities of Southeast Asia's rarest geckos: Flap-legged geckos (Luperosaurus), Flying geckos (Ptychozoon) and their relationship to the pan-Asian genus Gekko. Mol Phylogenet Evol 63:915-921.

Cannon JT, Vellutini BC, Smith J, 3rd, Ronquist F, Jondelius U, Hejnol A. 2016a. Xenacoelomorpha is the sister group to Nephrozoa. Nature 530:89-93.

Cannon JT, Vellutini BC, Smith J, Ronquist F, Jondelius U, Hejnol A. 2016b. Data from: Xenacoelomorpha is the sister group to Nephrozoa. In: Dryad Data Repository.

Chernomor O, von Haeseler A, Minh BQ. 2016. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. Syst. Biol. 65:997-1008.

Cognato AI, Vogler AP. 2001a. Data from: Exploring data interaction and nucleotide alignment in a multiple gene analysis of Ips (Coleoptera: Scolytinae). In: Dryad Data Repository.

Cognato AI, Vogler AP. 2001b. Exploring data interaction and nucleotide alignment in a multiple gene analysis of Ips (Coleoptera: Scolytinae). Syst. Biol. 50:758-780.

Day JJ, Peart CR, Brown KJ, Bills R, Friel JP, Moritz T. 2013. Data from: Continental diversification of an African catfish radiation (Mochokidae: Synodontis). In: Dryad Data Repository.

Day JJ, Peart CR, Brown KJ, Friel JP, Bills R, Moritz T. 2013. Continental diversification of an African catfish radiation (Mochokidae: Synodontis). Syst. Biol. 62:351-365.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics 6:361.

Devitt TJ, Cameron Devitt SE, Hollingsworth BD, McGuire JA, Moritz C. 2013. Data from: Montane refugia predict population genetic structure in the Large-blotched Ensatina salamander. In: Dryad Data Repository.

Devitt TJ, Devitt SE, Hollingsworth BD, McGuire JA, Moritz C. 2013. Montane refugia predict population genetic structure in the Large-blotched Ensatina salamander. Mol. Ecol. 22:1650-1665.

Dornburg A, Moore JA, Webster R, Warren DL, Brandley MC, Iglesias TL, Wainwright PC, Near TJ. 2012a. Data from: Molecular phylogenetics of squirrelfishes and soldierfishes (Teleostei:Beryciformes: Holocentridae): reconciling more than 100 years of taxonomic confusion. In: Dryad Data Repository.

Dornburg A, Moore JA, Webster R, Warren DL, Brandley MC, Iglesias TL, Wainwright PC, Near TJ. 2012b. Molecular phylogenetics of squirrelfishes and soldierfishes (Teleostei: Beryciformes: Holocentridae): reconciling more than 100 years of taxonomic confusion. Mol Phylogenet Evol 65:727-738.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

Duchene DA, Duchene S, Ho SYW. 2017. New Statistical Criteria Detect Phylogenetic Bias Caused by Compositional Heterogeneity. Mol. Biol. Evol. 34:1529-1534.

Dunn M, Greenhill SJ, Levinson SC, Gray RD. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. Nature 473:79.

Dutheil J, Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. BMC Evol. Biol. 8:255.

Eisen JA. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res. 8:163-167.

Faircloth BC, Sorenson L, Santini F, Alfaro ME. 2013a. Data from: A phylogenomic perspective on the radiation of ray-finned fishes based upon targeted sequencing of ultraconserved elements (UCEs). In: Dryad Data Repository.

Faircloth BC, Sorenson L, Santini F, Alfaro ME. 2013b. A Phylogenomic Perspective on the Radiation of Ray-Finned Fishes Based upon Targeted Sequencing of Ultraconserved Elements (UCEs). PLoS One 8:e65923.

Farrell LE, Roman J, Sunquist ME. 2000. Dietary separation of sympatric carnivores identified by molecular analysis of scats. Mol. Ecol. 9:1583-1590.

Felsenstein J. 2004. Inferring phylogenies: Sinauer associates Sunderland, MA.

Fong JJ, Brown JM, Fujita MK, Boussau B. 2012a. Data from: A phylogenomic approach to vertebrate phylogeny supports a turtle-archosaur affinity and a possible paraphyletic Lissamphibia. In: Dryad Data Repository.

Fong JJ, Brown JM, Fujita MK, Boussau B. 2012b. A phylogenomic approach to vertebrate phylogeny supports a turtle-archosaur affinity and a possible paraphyletic lissamphibia. PLoS One 7:e48990.

Foster PG. 2004. Modeling compositional heterogeneity. Syst. Biol. 53:485-495.

Foster PG, Hickey DA. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J. Mol. Evol. 48:284-290.

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S. 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419:498.

Goldman N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182-198.

Goremykin V, Hellwig F. 2005. Evidence for the most basal split in land plants dividing bryophyte and tracheophyte lineages. Plant Syst. Evol. 254:93-103.

Graur D, Gouy M, Duret L. 1997. Evolutionary affinities of the order Perissodactyla and the phylogenetic status of the superordinal taxa Ungulata and Altungulata. Mol Phylogenet Evol 7:195-200.

Gray RD, Drummond AJ, Greenhill SJ. 2009. Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement. Science 323:479.

Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, Holmes EC. 2004. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. Science 303:327.

Groussin M, Boussau B, Gouy M. 2013. A branch-heterogeneous model of protein evolution for efficient inference of ancestral sequences. Syst. Biol. 62:523-538.

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

Hasegawa M, Kishino H, Yano T-a. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160-174.

Ho JW, Adams CE, Lew JB, Matthews TJ, Ng CC, Shahabi-Sirjani A, Tan LH, Zhao Y, Easteal S, Wilson SR. 2006. SeqVis: visualization of compositional heterogeneity in large alignments of nucleotides. Bioinformatics 22:2162-2163.

Ho SY, Jermiin L. 2004. Tracing the decay of the historical signal in biological sequence data. Syst. Biol. 53:623-637.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. 35:518-522.

Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, Huelsenbeck JP, Ronquist F. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst. Biol. 65:726-736.

Horn JW, Xi Z, Riina R, Peirson JA, Yang Y, Dorsey BL, Berry PE, Davis CC, Wurdack KJ. 2014a. Data from: Evolutionary bursts in Euphorbia (Euphorbiaceae) are linked with photosynthetic pathway. In: Dryad Data Repository.

Horn JW, Xi Z, Riina R, Peirson JA, Yang Y, Dorsey BL, Berry PE, Davis CC, Wurdack KJ. 2014b. Evolutionary bursts in Euphorbia (Euphorbiaceae) are linked with photosynthetic pathway. Evolution 68:3485-3504.

Hyman IT, Ho SY, Jermiin LS. 2007. Molecular phylogeny of Australian Helicarionidae, Euconulidae and related groups (Gastropoda: Pulmonata: Stylommatophora) based on mitochondrial DNA. Mol. Phylogen. Evol. 45:792-812.

Jayaswal V, Ababneh F, Jermiin LS, Robinson J. 2011. Reducing Model Complexity of the General Markov Model of Evolution. Mol. Biol. Evol. 28:3045-3059.

Jayaswal V, Robinson J, Jermiin L. 2007. Estimation of Phylogeny and Invariant Sites under the General Markov Model of Nucleotide Sequence Evolution. Syst. Biol. 56:155-162.

Jayaswal V, Wong TK, Robinson J, Poladian L, Jermiin LS. 2014. Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. Syst. Biol. 63:726-742.

Jermiin L, Ho SY, Ababneh F, Robinson J, Larkum AW. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst. Biol. 53:638-643.

Jermiin LS, Jayaswal V, Ababneh FM, Robinson J. 2017. Identifying Optimal Models of Evolution. In: Keith JM, editor. Bioinformatics. Melbourne: Humana Press, New York, NY. p. 379-420.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14:587-589.

Kawahara AY, Rubinoff D. 2013a. Convergent evolution of morphology and habitat use in the explosive Hawaiian fancy case caterpillar radiation. J. Evol. Biol. 26:1763-1773.

Kawahara AY, Rubinoff D. 2013b. Data from: Convergent evolution in the explosive Hawaiian Fancy Cased caterpillar radiation. In: Dryad Data Repository.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111-120.

Kishino H, Miyata T, Hasegawa M. (Kishino1990 co-authors). 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. 31:151-160.

Knight R, Maxwell P, Birmingham A, Carnes J, Caporaso JG, Easton BC, Eaton M, Hamady M, Lindsay H, Liu Z. 2007. PyCogent: a toolkit for making sense from sequence. Genome biology 8:R171.

Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K. 2012. Statistics and truth in phylogenomics. Mol. Biol. Evol. 29:457-472.

Kumar S, Gadagkar SR. 2001. Disparity index: a simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. Genetics 158:1321-1327.

Lartillot N, Delsuc F. 2012a. Data from: Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. In: Dryad Data Repository.

Lartillot N, Delsuc F. 2012b. Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. Evolution 66:1773-1787.

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.

Liu L, Zhang J, Rheindt FE, Lei F, Qu Y, Wang Y, Zhang Y, Sullivan C, Nie W, Wang J, et al. 2017. Genomic evidence reveals a radiation of placental mammals uninterrupted by the KPg boundary. Proc Natl Acad Sci U S A 114:E7282-E7290.

Martijn J, Vosseberg J, Guy L, Offre P, Ettema TJ. 2018. Deep mitochondrial origin outside the sampled alphaproteobacteria. Nature.

Mäser P, Thomine S, Schroeder JI, Ward JM, Hirschi K, Sze H, Talke IN, Amtmann A, Maathuis FJM, Sanders D, et al. 2001. Phylogenetic Relationships within Cation Transporter Families of Arabidopsis. Plant Physiol. 126:1646.

McCormack JE, Harvey MG, Faircloth BC, Crawford NG, Glenn TC, Brumfield RT. 2013a. Data from: A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. In: Dryad Data Repository.

McCormack JE, Harvey MG, Faircloth BC, Crawford NG, Glenn TC, Brumfield RT. 2013b. A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. PLoS One 8:e54848.

Mir A, Russello F. 2010. The mean value of the squared path-difference distance for rooted phylogenetic trees. J Math Anal Appl 371:168-176.

Moyle RG, Oliveros CH, Andersen MJ, Hosner PA, Benz BW, Manthey JD, Travers SL, Brown RM, Faircloth BC. 2016a. Data from: Tectonic collision and uplift of Wallacea triggered the global songbird radiation. In: Dryad Data Repository.

Moyle RG, Oliveros CH, Andersen MJ, Hosner PA, Benz BW, Manthey JD, Travers SL, Brown RM, Faircloth BC. 2016b. Tectonic collision and uplift of Wallacea triggered the global songbird radiation. Nat Commun 7:12709.

Murray EA, Carmichael AE, Heraty JM. 2013a. Ancient host shifts followed by host conservatism in a group of ant parasitoids. Proc Biol Sci 280:20130495.

Murray EA, Carmichael AE, Heraty JM. 2013b. Data from: Ancient host shifts followed by host conservatism in a group of ant parasitoids. In: Dryad Data Repository.

Murray S, Jørgensen MF, Ho SY, Patterson DJ, Jermiin LS. 2005. Improving the analysis of dinoflagellate phylogeny based on rDNA. Protist 156:269-286.

Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ. 2001. Molecular phylogenetics and the origins of placental mammals. Nature 409:614-618.

Nabholz B, Künstner A, Wang R, Jarvis ED, Ellegren H. 2011. Dynamic Evolution of Base Composition: Causes and Consequences in Avian Phylogenomics. Mol. Biol. Evol. 28:2197-2210.

Nakagawa S, Schielzeth H. 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods in Ecology and Evolution 4:133-142.

Nesnidal MP, Helmkampf M, Bruchhaus I, Hausdorf B. 2010. Compositional Heterogeneity and Phylogenomic Inference of Metazoan Relationships. Mol. Biol. Evol. 27:2095-2104.

Nguyen LT, 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.

Oaks JR. 2011a. Data from: A time-calibrated species tree of Crocodylia reveals a recent radiation of the true crocodiles. In: Dryad Data Repository.

Oaks JR. 2011b. A time-calibrated species tree of Crocodylia reveals a recent radiation of the true crocodiles. Evolution 65:3285-3297.

Paton T, Haddrath O, Baker AJ. 2002. Complete mitochondrial DNA genome sequences show that modern birds are not descended from transitional shorebirds. Proceedings of the Royal Society of London B: Biological Sciences 269:839-846.

Philippe H, Brinkmann H, Copley RR, Moroz LL, Nakano H, Poustka AJ, Wallberg A, Peterson KJ, Telford MJ. 2011. Acoelomorph flatworms are deuterostomes related to Xenoturbella. Nature 470:255.

Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005. Phylogenomics. Annu Rev Ecol Evol S 36:541-562.

Phillips MJ, Delsuc Fdr, Penny D. 2004. Genome-Scale Phylogeny and the Detection of Systematic Biases. Mol. Biol. Evol. 21:1455-1458.

Rightmyer MG, Griswold T, Brady SG. 2013a. Data from: Phylogeny and systematics of the bee genus Osmia (Hymenoptera: Megachilidae) with emphasis on North American Melanosmia: subgenera, synonymies, and nesting biology revisited. In: Dryad Data Repository.

Rightmyer MG, Griswold T, Brady SG. 2013b. Phylogeny and systematics of the bee genus Osmia (Hymenoptera: Megachilidae) with emphasis on North American Melanosmia: subgenera, synonymies and nesting biology revisited. Syst. Entomol. 38:561-576.

Roberts D, Yang Z. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol. Biol. Evol. 12:451-458.

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

Rzhetsky A, Nei M. 1995. Tests of applicability of several substitution models for DNA sequence data. Mol. Biol. Evol. 12:131-151.

Salipante SJ, Horwitz MS. 2006. Phylogenetic fate mapping. Proceedings of the National Academy of Sciences 103:5448.

Sand A, Pedersen CNS, Brodal GS, Johansen J, Holt MK, Mailund T. 2014. tqDist: a library for computing the quartet and triplet distances between binary or general trees. Bioinformatics 30:2079-2080.

Sauquet H, Ho SY, Gandolfo MA, Jordan GJ, Wilf P, Cantrill DJ, Bayly MJ, Bromham L, Brown GK, Carpenter RJ, 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.

Sauquet H, Ho SYW, Gandolfo MA, Jordan GJ, Wilf P, Cantrill DJ, Bayly MJ, Bromham L, Brown GK, Carpenter RJ, et al. 2011. Data from: Testing the impact of calibration on molecular divergence times using a fossil-rich group: the case of Nothofagus (Fagales). In: Dryad Data Repository.

Seago AE, Giorgi JA, Li J, Ślipiński A. 2011a. Data from: Phylogeny, classification and evolution of ladybird beetles (Coleoptera: Coccinellidae) based on simultaneous analysis of molecular and morphological data. In: Dryad Data Repository.

Seago AE, Giorgi JA, Li J, Ślipiński A. 2011b. Phylogeny, classification and evolution of ladybird beetles (Coleoptera: Coccinellidae) based on simultaneous analysis of molecular and morphological data. Mol. Phylogen. Evol. 60:137-151.

Sharanowski BJ, Dowling APG, Sharkey MJ. 2011a. Data from: Molecular phylogenetics of Braconidae (Hymenoptera: Ichneumonoidea) based on multiple nuclear genes and implications for classification. In: Dryad Data Repository.

Sharanowski BJ, Dowling APG, Sharkey MJ. 2011b. Molecular phylogenetics of Braconidae (Hymenoptera: Ichneumonoidea), based on multiple nuclear genes, and implications for classification. Syst. Entomol. 36:549-572.

Sheffield NC, Song H, Cameron SL, Whiting MF. 2009. Nonstationary Evolution and Compositional Heterogeneity in Beetle Mitochondrial Phylogenomics. Syst. Biol. 58:381-394.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492-508.

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114-1116.

Siler C, Brown RM, Oliveros CH, Santanen A. 2013. Data from: Multilocus phylogeny reveals unexpected diversification patterns in Asian Wolf Snakes (genus Lycodon). In: Dryad Data Repository.

Siler CD, Oliveros CH, Santanen A, Brown RM. 2013. Multilocus phylogeny reveals unexpected diversification patterns in Asian wolf snakes (genus Lycodon). Zool. Scr. 42:262-277.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312-1313.

Steel MA, Penny D. 1993. Distributions of Tree Comparison Metrics - Some New Results. Syst. Biol. 42:126-141.

Stuart A. 1955. A Test for Homogeneity of the Marginal Distributions in a Two-Way Classification. Biometrika 42:412-416.

Sullivan J, Joyce P. 2005. Model selection in phylogenetics. Annual Review of Ecology Evolution and Systematics 36:445-466.

Sumner JG, Fernandez-Sanchez J, Jarvis PD. 2012. Lie Markov models. J. Theor. Biol. 298:16-31.

Swofford DL. 2001. Paup*: Phylogenetic analysis using parsimony (and other methods) 4.0. B5.

Tarrío R, Rodríguez-Trelles F, Ayala FJ. 2001. Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae. Mol. Biol. Evol. 18:1464-1473.

Tolley KA, Townsend TM, Vences M. 2013a. Data from: Large-scale phylogeny of chameleons suggests African origins and Eocene diversification. In: Dryad Data Repository.

Tolley KA, Townsend TM, Vences M. 2013b. Large-scale phylogeny of chameleons suggests African origins and Eocene diversification. Proc Biol Sci 280:20130184.

Tsagkogeorga G, Parker J, Stupka E, Cotton JA, Rossiter SJ. 2013. Phylogenomic analyses elucidate the evolutionary relationships of bats. Curr. Biol. 23:2262-2267.

Unmack PJ, Allen GR, Johnson JB. 2013a. Data from: Phylogeny and biogeography of rainbowfishes (Melanotaeniidae) from Australia and New Guinea. In: Dryad Data Repository.

Unmack PJ, Allen GR, Johnson JB. 2013b. Phylogeny and biogeography of rainbowfishes (Melanotaeniidae) from Australia and New Guinea. Mol Phylogenet Evol 67:15-27.

Wainwright PC, Smith WL, Price SA, Tang KL, Sparks JS, Ferry LA, Kuhn KL, Eytan RI, Near TJ. 2012. The evolution of pharyngognathy: a phylogenetic and functional appraisal of the pharyngeal jaw key innovation in labroid fishes and beyond. Syst. Biol. 61:1001-1027.

Wainwright PC, Smith WL, Price SA, Tang KL, Sparks JS, Ferry LA, Kuhn KL, Near TJ. 2012. Data from: The evolution of pharyngognathy: a phylogenetic and functional appraisal of the pharyngeal jaw key innovation in labroid fishes and beyond. In: Dryad Data Repository.

Weiss G, von Haeseler A. 2003. Testing Substitution Models Within a Phylogenetic Tree. Mol. Biol. Evol. 20:572-578.

Wood HM, Matzke NJ, Gillespie RG, Griswold CE. 2012. Data from: Treating fossils as terminal taxa in divergence time estimation reveals ancient vicariance patterns in the palpimanoid spiders. In: Dryad Data Repository.

Wood HM, Matzke NJ, Gillespie RG, Griswold CE. 2013. Treating fossils as terminal taxa in divergence time estimation reveals ancient vicariance patterns in the palpimanoid spiders. Syst. Biol. 62:264-284.

Woodhams MD, Fernandez-Sanchez J, Sumner JG. 2015. A New Hierarchy of Phylogenetic Models Consistent with Heterogeneous Substitution Rates. Syst. Biol. 64:638-650.

Worobey M, Han G, Rambaut A. 2014a. Data from: A synchronized global sweep of the internal genes of modern avian influenza virus. In: Dryad Data Repository.

Worobey M, Han GZ, Rambaut A. 2014b. A synchronized global sweep of the internal genes of modern avian influenza virus. Nature 508:254-257.

Yang Z. 1994. Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39:105-111.

Yang Z, Rannala B. 2012. Molecular phylogenetics: principles and practice. Nat. Rev. Genet. 13:303-314.

Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, Kimmel M, Kavraki L, Lichtarge O. 2003. An accurate, sensitive, and scalable method to identify functional sites in protein structures. J. Mol. Biol. 326:255-261.

Yao Y-G, Bravi CM, Bandelt H-J. 2004. A call for mtDNA data quality control in forensic science. Forensic Sci. Int. 141:1-6.

Yap VB, Speed T. 2005. Rooting a phylogenetic tree with nonreversible substitution models. BMC Evol. Biol. 5:2.

Zhou X, Xu S, Xu J, Chen B, Zhou K, Yang G. 2012. Phylogenomic analysis resolves the interordinal relationships and rapid diversification of the laurasiatherian mammals. Syst. Biol. 61:150-164.

Zou L, Susko E, Field C, Roger AJ. 2012. Fitting nonstationary general-time-reversible models to obtain edge-lengths and frequencies for the Barry–Hartigan model. Syst. Biol. 61:927-940.

Supplementary Data

Extended Table 1| best-fitting model by ModelFinder and number of partitions that got each model as the best-fit model. Finding the best-fitting model (which minimize BIC score) for each one of the partitions.

Substitution model	#partitions with best-fit model	Nucleotide frequencies
K80	667	Equal
HKY	563	Unequal
TPM2u	297	Unequal
TPM3u	291	Unequal
GTR	272	Unequal
TVM	234	Unequal
TIM3	164	Unequal
TIM2	124	Unequal
TN	110	Unequal
TVMe	106	Equal
TNe	98	Equal
TIM3e	84	Equal
K81	83	Equal
K81u	82	Unequal
TIM2e	80	Equal
TPM3	60	Equal
TPM2	54	Equal
SYM	53	Equal
TIM	41	Unequal
JC	32	Equal
F81	21	Unequal
TIMe	16	Equal

Extended Table 2| best-fitting model by ModelFinder for the partitions that passed each one of the three max-value tests.

MaxSymTest	
model	#partitions
K80	548
HKY	514
GTR	256
TPM2u	243
TPM3u	239
TVM	177
TIM3	135
TIM2	102
TN	95
TNe	76
K81u	66
TPM3	65
TVMe	65
TIM3e	64
K81	64
TIM2e	56
TPM2	49
TIM	34
JC	34
SYM	33
F81	20
TIMe	12

MaxSymTest$_{mar}$	
model	#partitions
K80	503
HKY	489
GTR	238
TPM3u	236
TPM2u	232
TVM	163
TIM3	138
TIM2	90
TN	84
TNe	76
K81	63
TIM3e	61
TVMe	59
TPM3	56
K81u	55
TIM2e	51
TPM2	42
SYM	33
TIM	32
JC	29
F81	18
TIMe	13

MaxSymTest$_{int}$	
model	#partitions
K80	622
HKY	522
TPM3u	287
TPM2u	287
GTR	266
TVM	224
TIM3	151
TIM2	123
TVMe	100
TN	96
TNe	91
K81	81
TIM3e	77
K81u	75
TIM2e	74
TPM3	57
SYM	51
TPM2	51
TIM	39
JC	28
TIMe	17
F81	15

Extended Table 3| best-fitting model by ModelFinder for the partitions that failed each one of the three max-value tests.

MaxSymTest	
model	#partitions
K80	128
HKY	59
TVM	48
TPM2u	46
TPM3u	38
TVMe	33
GTR	27
TIM2e	25
TNe	21
TIM3	21
TIM2	21
K81	20
TN	18
TIM3e	14
SYM	13
K81u	12
TIM	7
TPM2	7
TIMe	6
F81	2
JC	0
TIMe	0

MaxSymTest$_{mar}$	
model	#partitions
K80	157
HKY	74
TVM	68
TPM2u	52
TPM3u	51
GTR	41
TVMe	36
TIM2	33
TIM2e	27
TNe	23
TIM3e	21
TN	21
TIM3	21
K81u	21
K81	21
SYM	17
TPM3	15
TPM2	10
TIM	8
TIMe	6
JC	3
F81	2

MaxSymTest$_{int}$	
model	#partitions
HKY	22
K80	14
TVM	9
GTR	8
TIM3	8
TN	7
TPM3u	5
TPM2	4
TVMe	4
TIM2	4
SYM	3
JC	3
TIM2e	3
TPM3	2
TIM	1
TIM3e	1
K81u	1
TPM2u	1
K81	1
SYM	0
F81	0
TIMe	0

Extended Table 4| The quartet distances between the three trees (T_{all}, T_{pass}, T_{fail}) in MaxSymTest, MaxSymTest$_{mar}$, and MaxSymTest$_{int}$.

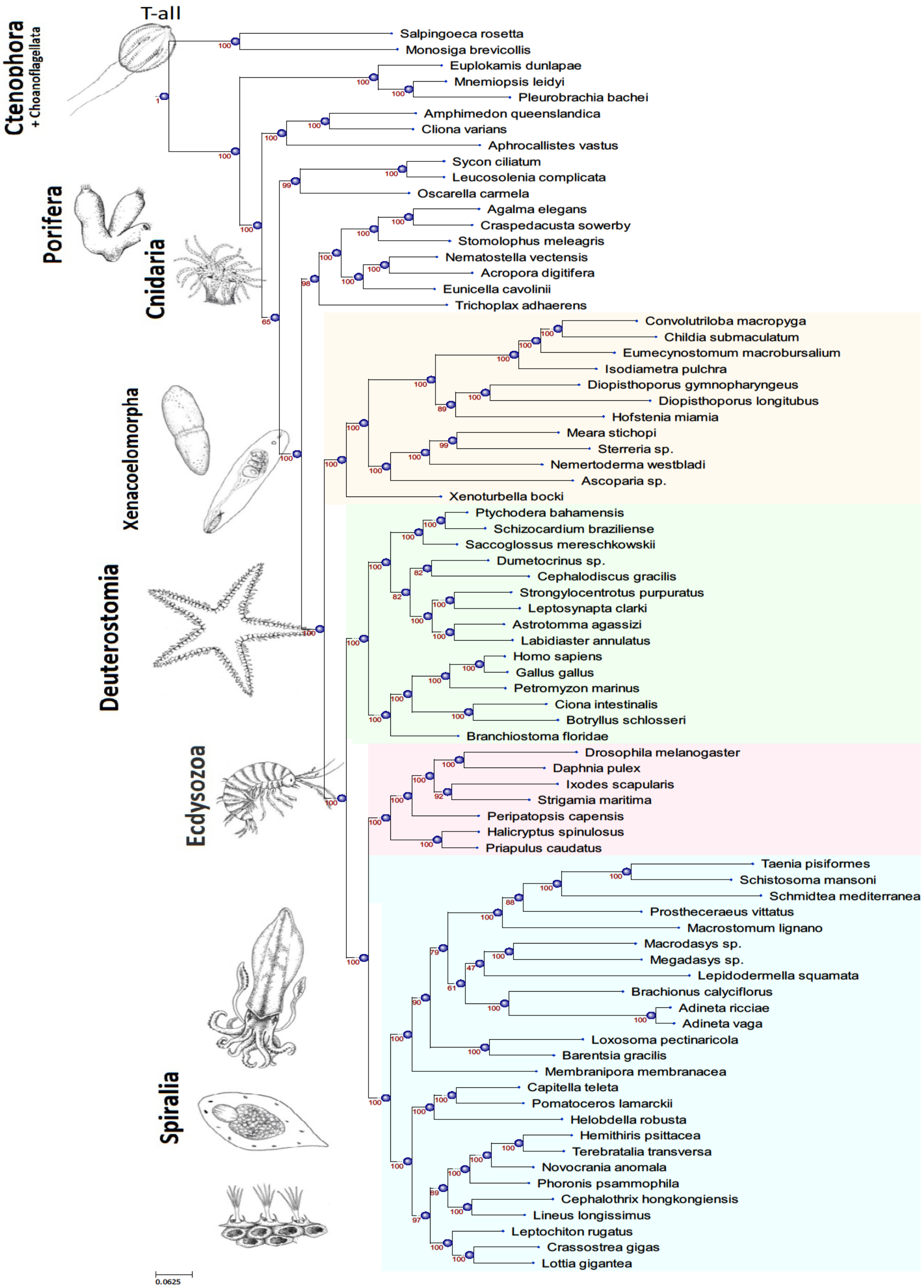
	Dataset	$T_{all-fail}$	$T_{all-pass}$	$T_{fail-pass}$
MaxSymTest	Anderson_2013	81372	4440183	4505628
	Bergsten_2013	25492	3902	27430
	Broughton_2013	35360	3738	39098
	Cannon_2016a	20809	5746	26555
	Dornburg_2012	2373	6992	9365
	Faircloth_2013	442	0	442
	Horn_2014	1180177	975250	1823719
	Kawahara_2013	95727	38539	132150
	Lartillot_2012	303589	18326	297248
	McCormack_2013	10195	1243	10749
	Moyle_2016	78998	3031	82029
	Oaks_2011	68452	4142	72582
	Rightmyer_2013	441077	183468	568615
	Siler_2013	29961	11064	32949
	Wainwright_2012	3897669	1546897	5368384
	Wood_2012	539	13135	13076
	Worobey_2014a	2699881	336202	2823981
	Worobey_2014c	627673	8616	631824
	Worobey_2014e	12707531	145471345	148579172
	Worobey_2014f	160855578	4754274	162732984
	Worobey_2014g	428217	22909429	23329085
	Worobey_2014h	248010	43931488	44066427
MaxSymTest$_{mar}$	Anderson_2013	130354	2284760	2294028
	Bergsten_2013	25492	8061	28962
	Broughton_2013	35360	3680	39040
	Cannon_2016a	24582	3163	27745
	Day_2013	1458532	3555949	4668546
	Dornburg_2012	4052	16758	18728
	Faircloth_2013	442	0	442
	Horn_2014	1337822	364531	1365852
	Kawahara_2013	123280	15164	129332
	Lartillot_2012	21740	24730	43494
	McCormack_2013	6156	2674	7154
	Moyle_2016	90698	3031	93729
	Oaks_2011	83990	4004	87402
	Rightmyer_2013	426883	355486	595726
	Siler_2013	30018	10778	32835

	Wainwright_2012	4291174	2964931	5995967
	Wood_2012	58	11167	11223
	Worobey_2014a	2688767	336156	2796992
	Worobey_2014b	4846696	178703912	179250141
	Worobey_2014c	268995	8583	275550
	Worobey_2014d	21784654	92744994	78398319
	Worobey_2014e	22662896	159964092	170213975
	Worobey_2014f	3293336	14028740	17125648
	Worobey_2014g	428217	22806396	23226052
	Worobey_2014h	7666956	24536879	31423451

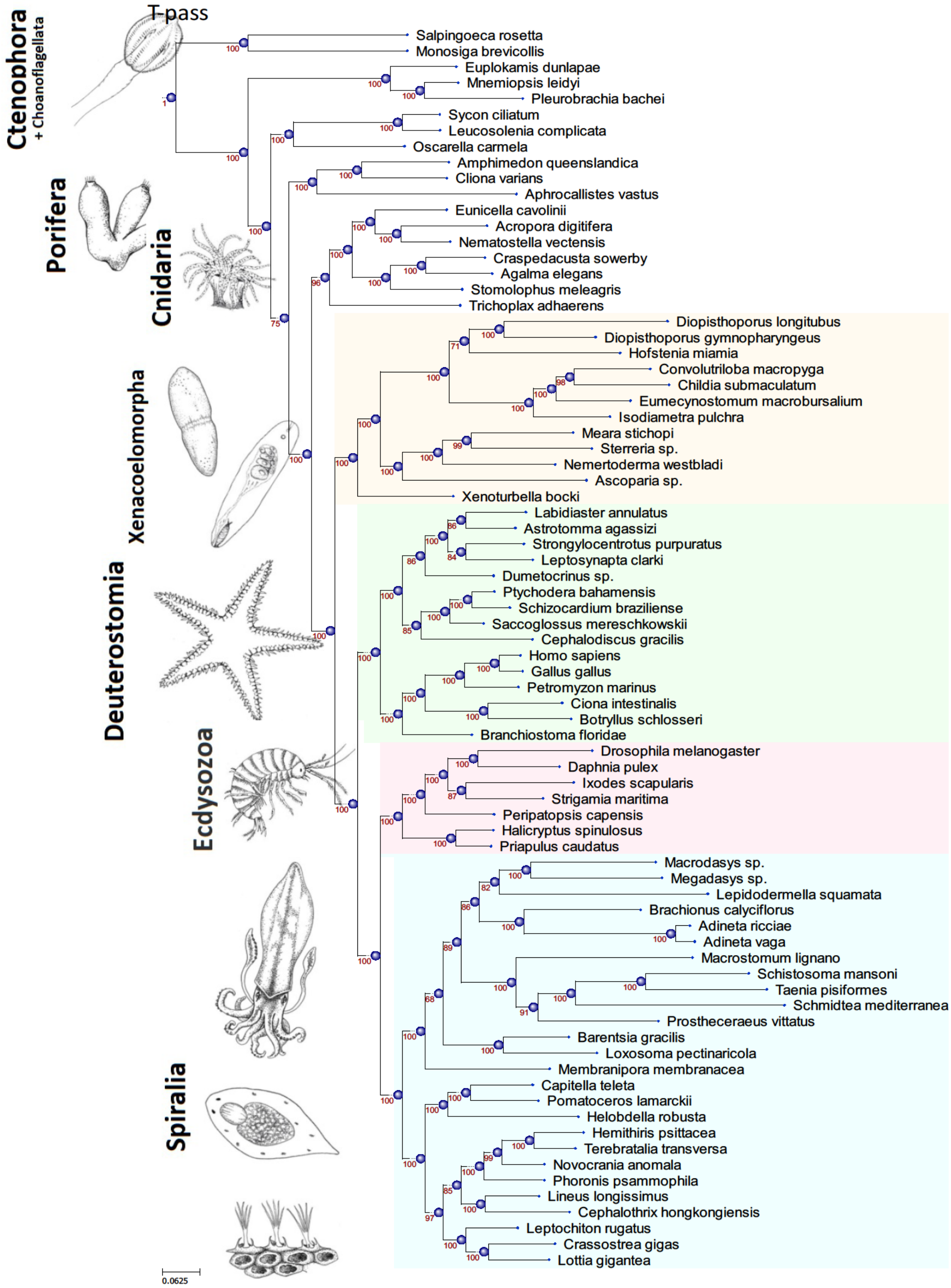
MaxSymTest$_{int}$	Cannon_2016a	221942	1323	223265
	Cognato_2001	1769	1728	41
	Faircloth_2013	412	0	412
	McCormack_2013	14090	687	14063
	Moyle_2016	75966	26979	102426
	Wood_2012	5867	3831	8679
	Worobey_2014f	134222118	2786430	135193257
	Worobey_2014h	7401114	22606490	27833436

Extended Table 5| Number datasets that contain loci from the different types of genomes and the number of partitions from each type of genome.

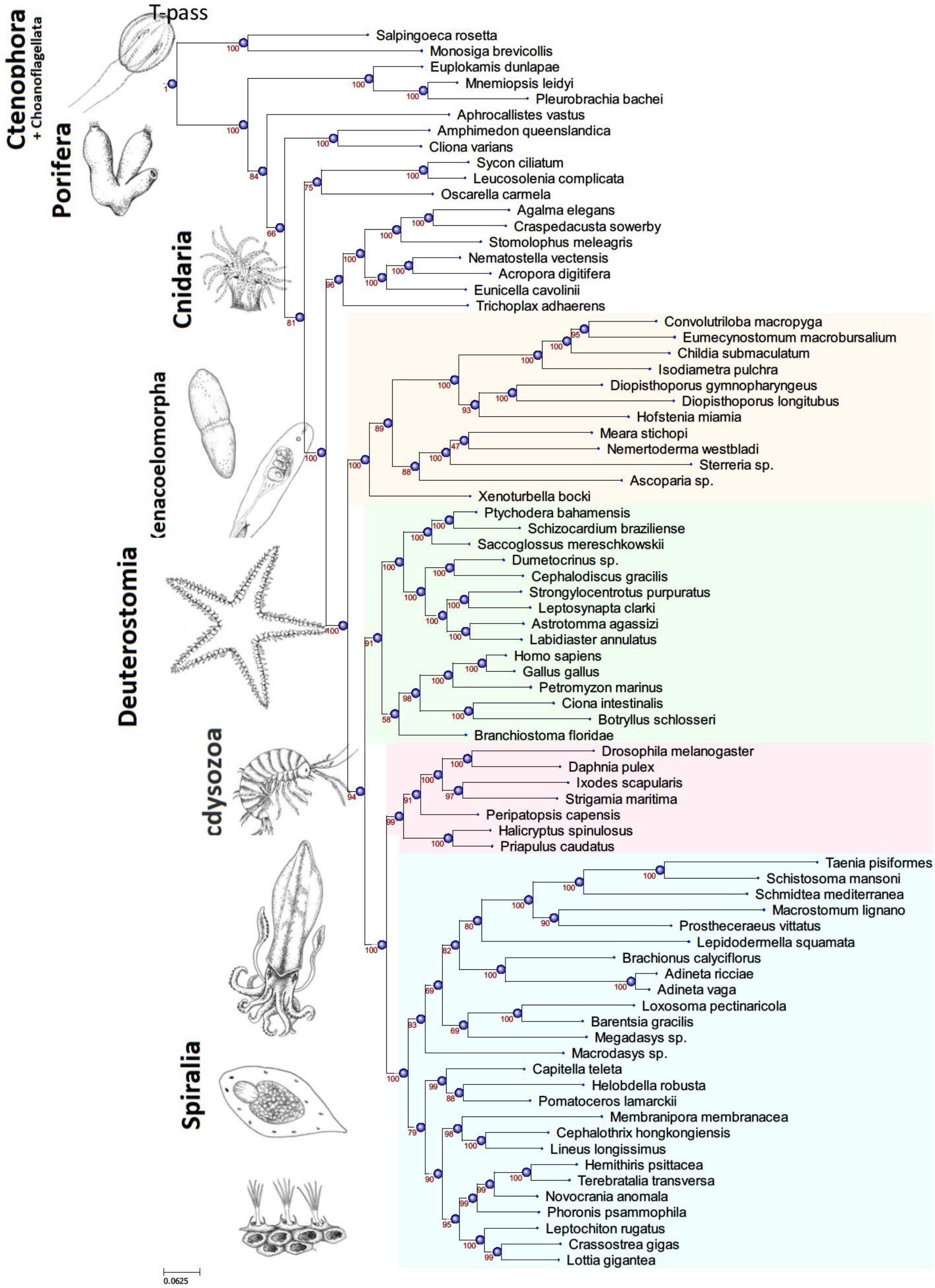
Genome type	#datasets	#genes	# partitions
Mitochondria	18	30	105
Nuclear	25	352	3419
Plastid	2	6	24
Virus	8	8	24



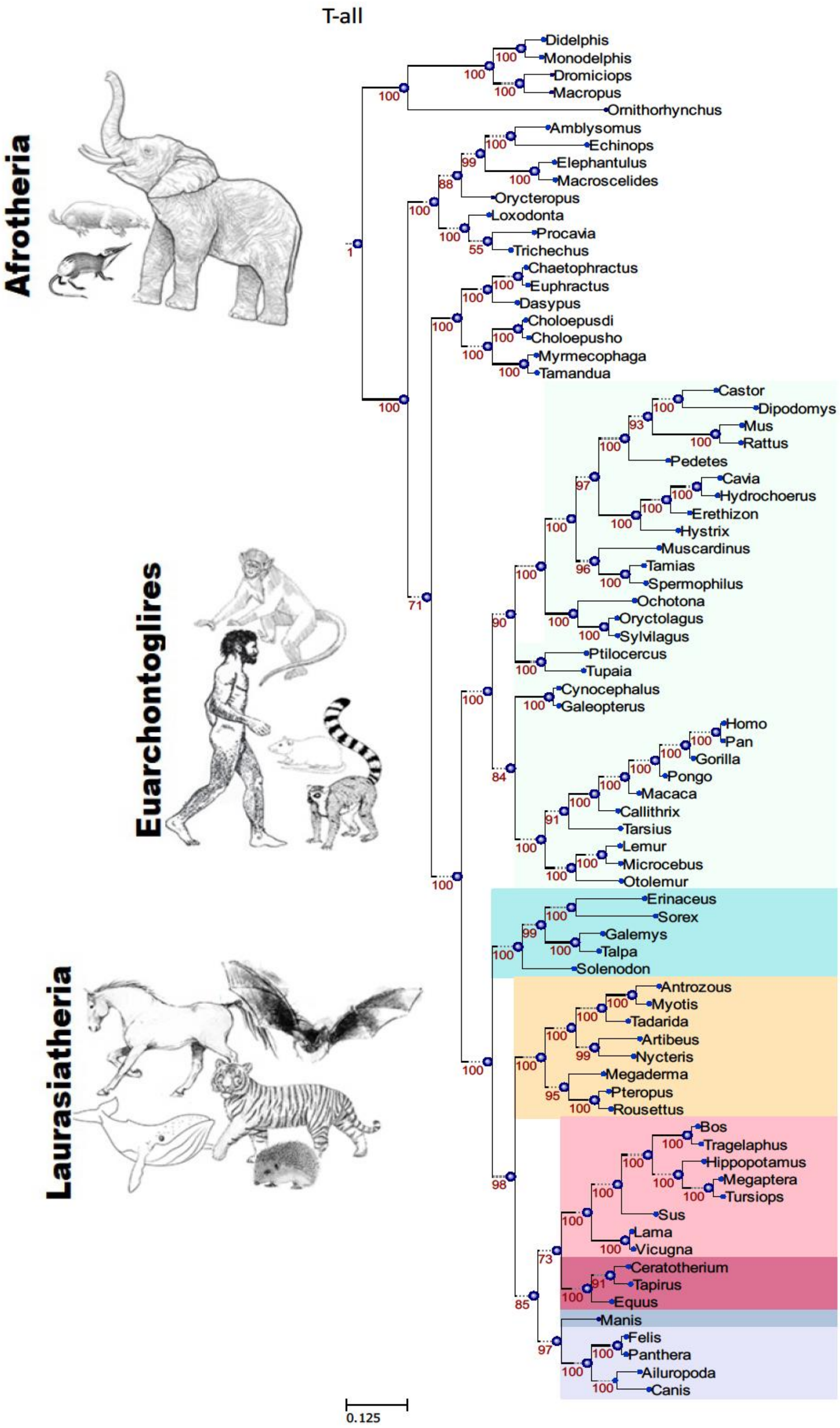
Extended Figure 1| ML topology of Cannon_2016 dataset inferred from all 424 partitions.



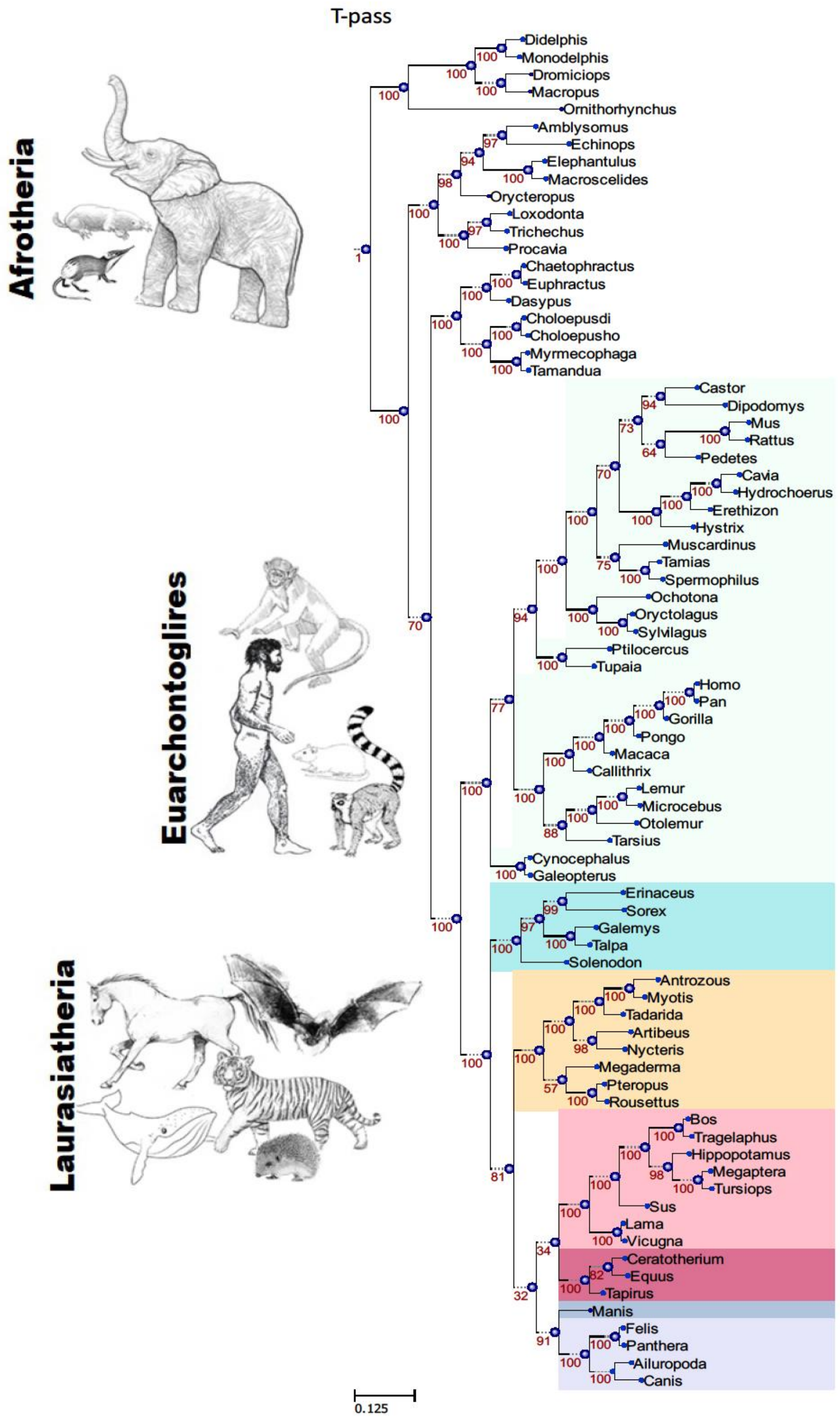
Extended Figure 2| ML topology of Cannon_2016 dataset inferred from all 281 partitions that passed the MaxSymTest.



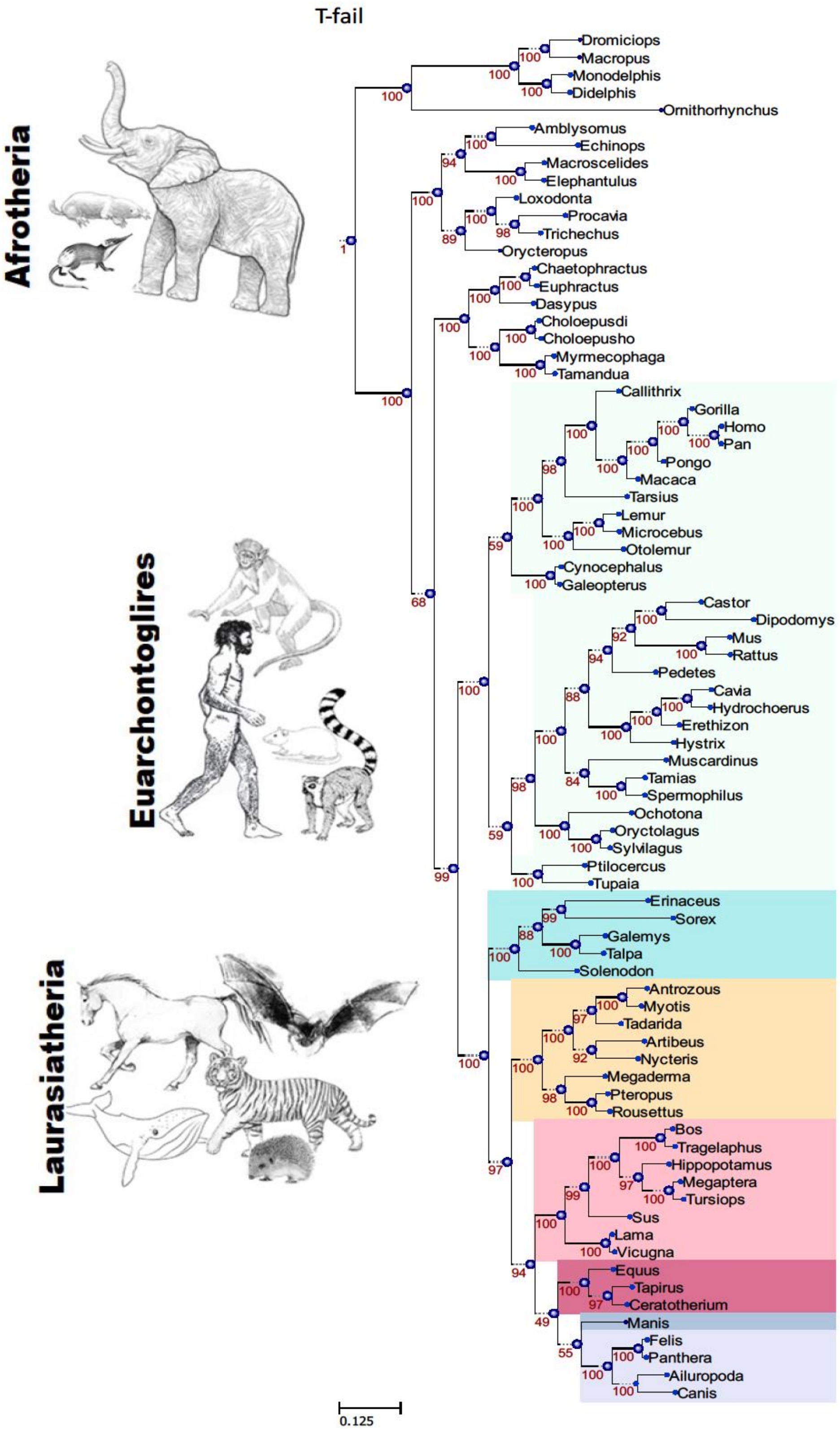
Extended Figure 3| ML topology of Cannon_2016 dataset inferred from all 143 partitions that failed the MaxSymTest.



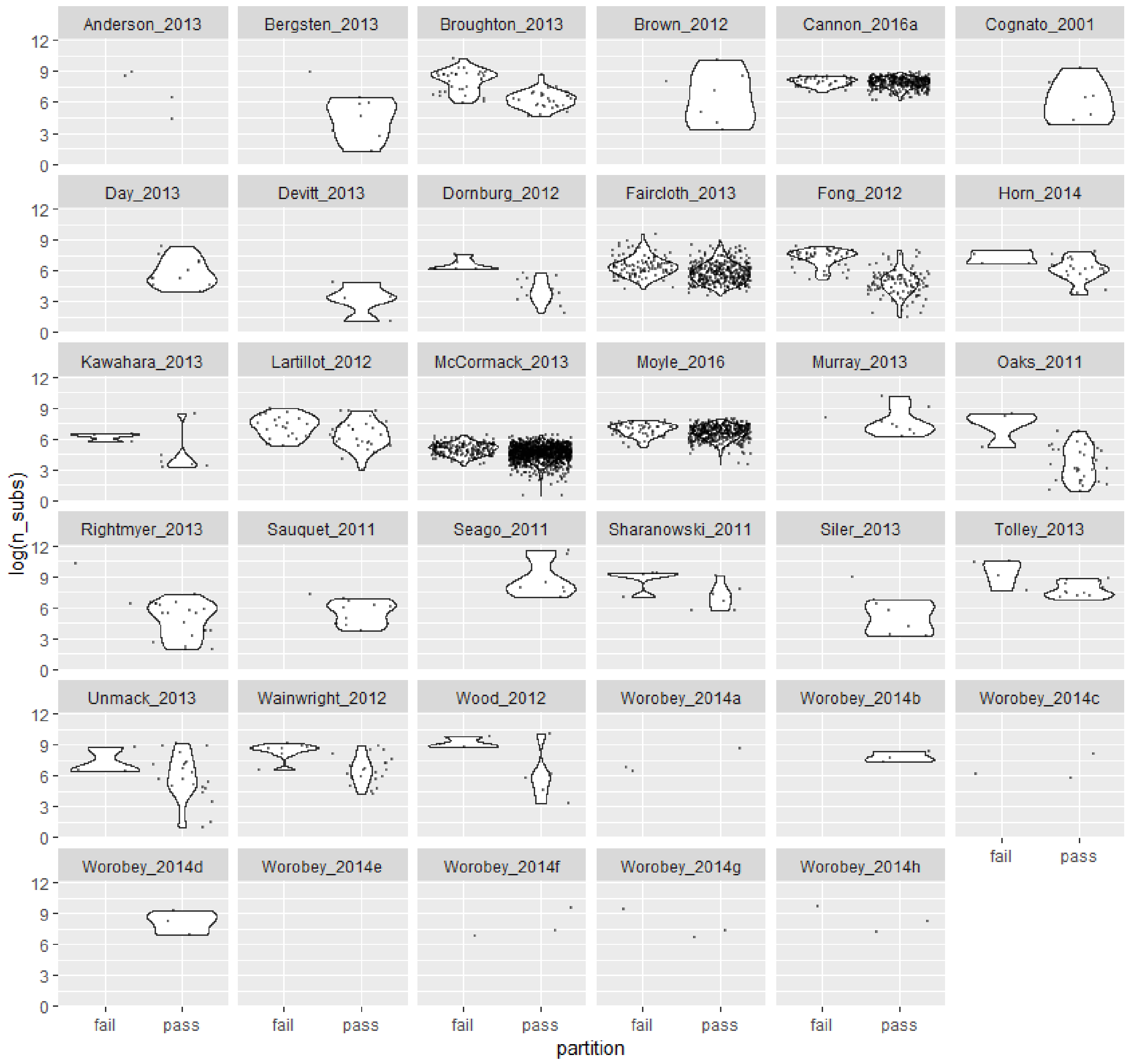
Extended Figure 4| ML topology of Lartillot_2012 dataset inferred from all 51 partitions.



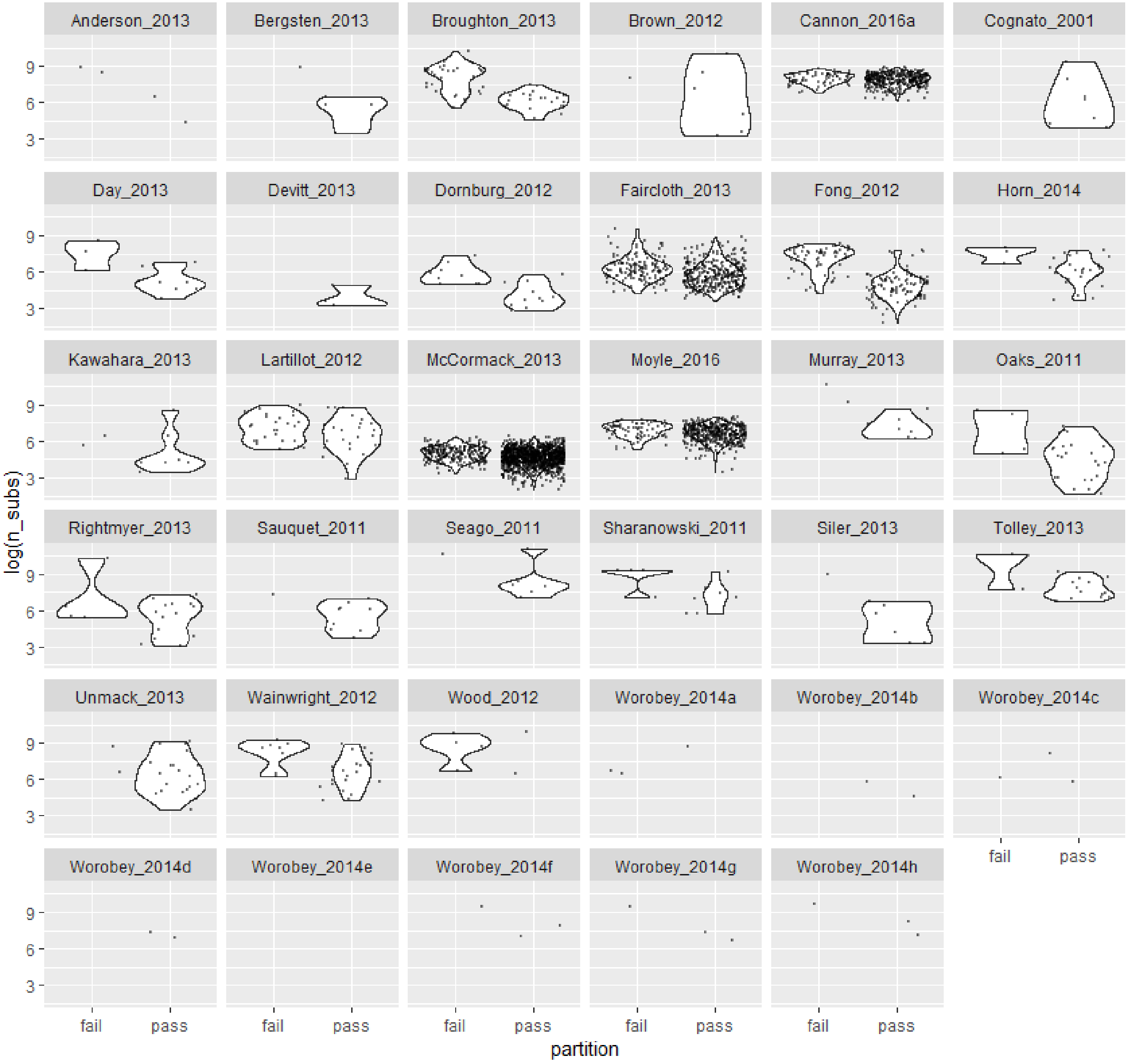
Extended Figure 5| ML topology of Lartillot_2012 dataset inferred from all 29 partitions that passed the MaxSymTest.



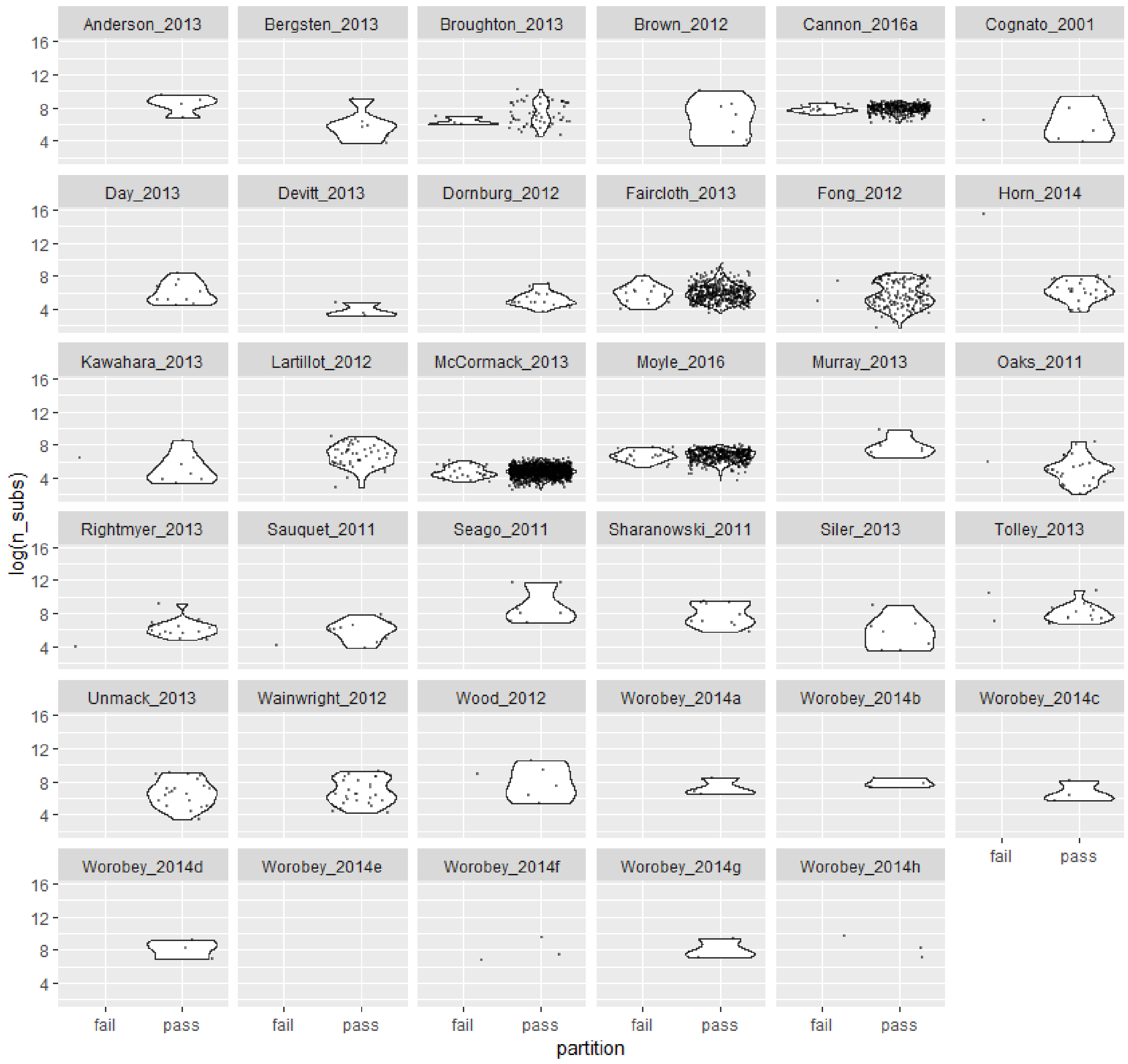
Extended Figure 6| ML topology of Lartillot_2012 dataset inferred from all 22 partitions that failed the MaxSymTest.



Extended Figure 7| The number of substitutions in partitions that failed or passed the MaxSymTest for each dataset.



Extended Figure 8| The number of substitutions in partitions that failed or passed the MaxSymTest$_{\text{mar}}$ for each dataset.



Extended Figure 9| The number of substitutions in partitions that failed or passed the MaxSymTest$_{\text{int}}$ for each dataset.

CHAPTER 2

THE INFLUENCE OF MODEL VIOLATION ON PHYLOGENETIC INFERENCE: A SIMULATION STUDY

Suha Naser-Khdour*[1], Bui Quang Minh[2], and Robert Lanfear[1]

(3)Department of Ecology and Evolution, Research School of Biology, Australian National University, Canberra, Australian Capital Territory, Australia

(4)School of Computing, Australian National University, Canberra, Australian Capital Territory, Australia

**Author for Correspondence: E-mail: suha.naser@anu.edu.au*

Contributions:

SNK wrote the python script, performed the analysis, analysed and interpreted the results, drafted the manuscript, and submitted the article for publication. MB contributed to the research design, conceptual development, and editorial comments. RL contributed to the research design, conceptual development and editorial comments.

Abstract

Phylogenetic inference typically assumes that the data has evolved under Stationary, Reversible and Homogeneous (SRH) conditions. Many empirical and simulation studies have shown that assuming SRH conditions can lead to significant errors in phylogenetic inference when the data violates these assumptions. Yet, many simulation studies focused on extreme non-SRH conditions that represent worst-case scenarios and not the average empirical dataset. In this study, we simulate datasets under various degrees of non-SRH conditions using empirically derived parameters to mimic real data and examine the effects of incorrectly assuming SRH conditions on inferring phylogenies. Our results show that maximum likelihood inference is generally quite robust to a wide range of SRH model violations but is inaccurate under extreme convergent evolution.

[Phylogenetic inference, model violations, systematic bias, simulations, evolution under non-SRH conditions, test of symmetry]

Main Text

Markov processes are commonly used in model-based phylogenetic analyses such as maximum likelihood (ML) and Bayesian inference (Felsenstein 2004; Yang 2006). A Markov model is represented by an instantaneous rate matrix Q of size 4-by-4 for DNA or 20-by-20 for protein sequences, that describes the substitution rates between nucleotides or amino acids (henceforth denoted as states), respectively. The Markovian propriety is convenient because the probabilities of the next states only depend on the current states, independently of how the current states had evolved (Felsenstein 1981; Felsenstein 1983; Yang 1994; Swofford, et al. 1996; Yang 2006). For mathematical simplicity and computational tractability, most studies assume that the Markov model is stationary, reversible, and homogeneous (SRH) (Kimura 1980; Felsenstein 1981; Hasegawa, et al. 1985; Tavaré 1986; Tamura and Nei 1993; Yang 1994). Homogeneity means that a single Q matrix operates along all edges of the tree, i.e., all substitution rates stay constant through time. Stationarity means that the state frequencies also remain constant along all edges of the tree. Reversibility means that the rate of change from state A to another state B is the same as the backward substitution rate from B to A.

The assumptions of homogeneity, stationarity, and reversibility come at the cost of complying with biological reality (Roberts and Yang 1995; Foster and Hickey 1999; Foster 2004; Ababneh, et al. 2006). For example, the reversibility assumption implies that the likelihood of a tree topology will be the same regardless of the placement of the root (Felsenstein 1981). Moreover, a reversible substitution model has up to 8 free rate parameters for nucleotides and 208 for amino acids, while a non-reversible substitution model has up to 11 free rate parameters for nucleotides and 379 for amino acids, provided that the model is still stationary and homogeneous (Yang 1994). These degrees of freedom can increase dramatically if the model is non-stationary or/and non-homogeneous (Barry and Hartigan 1987; Boussau and Gouy

2006): at the limit there can be an independent model of evolution on every branch of a tree, meaning that the total number of parameters is the product of the number of parameters in the substitution model and the number of branches in the tree.

Using stationary, reversible, and homogeneous substitution models to infer a phylogeny from data that has evolved under more complex conditions compromises the consistency of the ML estimation (Felsenstein 2004). Ideally, we would like to use data that comply with the assumptions of the models we apply, or alternatively, use models that are not violated by the data in hand. However, the use of non-SRH models is computationally demanding and is often not practical in large datasets. On the other hand, removing data that do not comply with the SRH assumption will come at a cost of losing phylogenetic information. Both simulation (Huelsenbeck and Hillis 1993; Hillis, et al. 1994; Galtier and Gouy 1998; Ho and Jermiin 2004; Jermiin, et al. 2004; Boussau and Gouy 2006) and empirical (Phillips, et al. 2004; Collins, et al. 2005; Nguyen, et al. 2012; Betancur, et al. 2013; Naser-Khdour, et al. 2019) studies have shown that applying SRH models to data that have evolved under more complex conditions can lead to significant errors in phylogenetic inference. However, most of these simulation studies have used parameters that do not reflect most empirical datasets, and sometimes represent extreme conditions such as the independent convergence of distantly-related taxa to a GC content that differs substantially from the rest of the taxa in the tree. While these simulations are based on biological observations such as the evolution of extreme GC content differences among closely related bacteria (Mooers and Holmes 2000), they do not represent the degree of violation of SRH conditions typical of most datasets. Indeed, apart from extreme cases it remains relatively poorly understood to what extent different types and degrees of violations of the SRH conditions affect phylogenetic inference.

In this study, we examine the influence of violating the SRH assumptions on phylogenetic inference with SRH models using parameters that are derived from thousands of empirical datasets. We simulate nucleotide alignments under various non-stationary (and thus non-reversible) or/and non-homogeneous conditions and examine the effects of incorrectly assuming SRH conditions on inferring phylogenies from these data. Moreover, we examine the ability of different methods to detect non-SRH evolution across multiple sequence alignments. Several tests for detecting non-SRH evolution in nucleotide and amino acid alignments have been introduced (Lanave, et al. 1984; Lanave, et al. 1986; von Haeseler, et al. 1993; Lockhart, et al. 1994; Kumar and Gadagkar 2001; Phillips and Penny 2003; Weiss and von Haeseler 2003; Foster 2004; Ababneh, et al. 2006; Ho, et al. 2006; Jermiin, et al. 2019; Naser-Khdour, et al. 2019). However, these tests are rarely used in phylogenetic analysis (Jermiin, et al. 2004; Jermiin, et al. 2009), likely because many of them are difficult to apply in practice. In this study, we focussed on three tests for detecting non-SRH evolution that are implemented in the widely-used IQ-TREE software (Minh, et al. 2020): the MaxSymTests (Naser-Khdour, et al. 2019), the compositional chi-square test (Preparata and Saccone 1987) as implemented in IQ-TREE (Nguyen, et al. 2015), and the test of non-stationarity proposed by Weiss and von Haeseler (Weiss and von Haeseler 2003). The MaxSymTests ask whether there is evidence in a single alignment that evolutionary symmetry imposed by SRH evolution is violated, and is a relatively new extension of similar tests designed for pairs of sequences (Jermiin, et al. 2019). The Weiss and von Haeseler (WH) test checks the homogeneity of the substitution model across the tree based on the pairwise sequence comparisons and performs a parametric bootstrap to assess the statistical significance (Weiss and von Haeseler 2003). The compositional chi-square test checks if the state composition of each sequence in the alignment is similar to the average state composition of the whole alignment, and is commonly-used to detect and sometimes remove sequences that clearly violate the SRH conditions (e.g. Aouad, et al. 2018; Liu, et al. 2018; Martijn, et al. 2018; Puttick, et al. 2018; Song, et al. 2018; Fan, et al. 2020). The Chi-square test gives researchers a way of understanding whether each sequence in an alignment has state frequencies that are plausible given the overall state frequencies of the alignment. We know of no existing test which combines individual chi-square tests to assess whether the state frequencies across all sequences of an alignment is plausible under an SRH model. It is possible to do this with model adequacy tests, but this requires one to first fit a full model and a tree (Foster 2004; Brown and ElDabaje 2009; Duchene, et al. 2017), while our current work focusses on tests that can be performed quickly and efficiently on very large datasets prior to tree inference. We therefore use two different approaches in this study to leverage the information in from individual chi-square tests.

The two approaches we take to using information from chi-square tests reflect different ways of balancing false-positive and false-negative outcomes, and so may be thought of as appropriate for different situations. Our first approach to using the chi-square tests is to take the most conservative possible approach and score an alignment as violating SRH assumptions if at least one sequence fails the test. Using the Chi-square frequencies in this way is very conservative, and liable to have a high false-positive rate that increases with the number of sequences in an alignment. However, in some practical cases when many loci are available but only a small number can be used for analyses, e.g. selecting ~50 loci for a Bayesian analysis out of many thousands available from whole genomes, a conservative approach such as this with a high false-positive rate may be warranted. Our second approach is less conservative. In this approach, we record the proportion of sequences in an alignment that fail the Chi-square test and ask whether this proportion is correlated with the degree of non-stationarity in the simulations. This approach may be more useful in practical cases where researchers wish to rank a set of loci with respect to the severity of model violations.

Materials and Methods

Simulations

In order to investigate the ability of SRH models to correctly infer topologies and branch lengths from non-SRH data, we devised a new approach that allows us to simulate alignments gradually ranging from true SRH conditions (with identical base frequencies and identical reversible substitution processes on every branch of the topology) to the most extreme violation with completely unrelated base frequencies and non-reversible substitution processes on every branch of the topology. For an alignment of m taxa and n sites, we will denote the set of all branches in the rooted tree τ as $\Phi = \{1, \dots, l\}$.

We simulate data under two different simulation schemes as follows:

1. An inheritance scheme designed to reflect the evolutionary process, in which each node in the tree inherits its substitution processes from its parent with a constant strength of inheritance modified by the branch length connecting the two nodes. The scheme reflects the continuity of evolutionary processes that are changing through time along a phylogenetic tree.
2. A two-matrix scheme designed to reflect previous approaches to simulating non-SRH evolution, where two independent subtrees (that are not sisters nor descendants of each other) have an identical substitution process and that is distinct from the substitution process that operates on the rest of the tree. This scheme resembles convergent evolution.

Applying these two schemes allows us to ask how evolutionarily-inspired non-SRH simulations are affected by SRH assumptions (scheme 1) and then to directly compare these to the more extreme forms of non-SRH evolution that are more often simulated (scheme 2). We will describe both simulation approaches in more detail below. But we start by describing how we choose model parameters for our simulations.

Estimating Empirical Parameter Distributions and Tree Topologies for Simulations

Both of our simulation approaches require us to choose base frequency vectors and rate matrices with which to simulate alignments. Generating these at random could limit the applicability of our results because it is unlikely that randomly-generated base frequency vectors or rate matrices would reflect reality. To address this, we instead estimated base frequency vectors and rate matrices from a large collection of empirical alignments, and then used these parameters for our simulations.

In order to estimate the distributions of the empirical base frequencies (Π) and the substitution rates (X) we used 32,666 partitions from 49 nucleotide datasets (Appendix Table A.1). Consisting of different types of partitions (codon positions, rRNA, tRNA, introns, intergenic spacers and UCEs) and genomes (nuclear, mitochondria, virus, plastid). Since different partitions of the genome evolve differently, for each partition, we ran IQ-TREE with a GTR model and free rate heterogeneity across sites (Yang 1995) with 4 categories + invariant sites. This gave us the distributions of 32,666 estimates of each parameter in the GTR matrix (A↔C, A↔G, A↔T, C↔G, C↔T, G↔T) and the distribution of each base frequency (π_A, π_C, π_G, π_T).

We use a similar approach to estimate the distribution of branch lengths. Estimating branch lengths from each partition separately could be misleading because there tends to be a high stochastic error in branch lengths estimated from short single-partition alignments (Kumar, et al. 2012). Therefore, in order to estimate the empirical distribution of the branch lengths, we instead estimated a single set of branch lengths from each of our 49 nucleotide datasets and complemented these with an additional 18 amino-acid datasets. For each dataset, we ran IQ-TREE with the best-fit fully-partitioned model (Chernomor, et al. 2016), which allows each partition to have its own evolutionary model and edge-linked rates determined by ModelFinder (Kalyaanamoorthy, et al. 2017). We then rooted the tree with the outgroup taxa (if provided) and extracted the empirical branch lengths of the ingroup (T) for each of the 33,178 partitions from 67 nucleotide and amino acid datasets.

Finally, for each parameter in X (5 parameters - G↔T equals to 1) and Π (4 parameters), and for each distribution in T (67 distributions - each dataset is an independent distribution) we find the best-fit distribution from 36 common probability distributions using the Kolmogorov-Smirnov test using SciPy (Virtanen, et al. 2020). We then sampled parameters for our simulations from these best-fit distributions. Since the parameters of Π are not independent, to sample a base-frequency vector we randomly sampled a parameter from each of the four base frequency's best-fit distribution and then normalized these parameters to sum to 1.

The tree topology τ is derived from birth-death simulations with speciation rate λ, extinction rate μ and the fraction of sampled taxa f using TreeSim package with a fixed number of extant species (Stadler 2011). In principle, it is possible to estimate the speciation and extinction rates from empirical data (Nee, et al. 1994; Rannala and Yang 1996; Magallon and Sanderson 2001). However, not knowing the fraction of sampled taxa a priori will tend to bias such estimates (Stadler 2013; Hua and Lanfear 2018). Because of the challenges of reliably estimating empirical speciation and extinction rates, we instead randomly sampled the speciation rate, the extinction rate and the fraction of sampled taxa from uniform distributions, to attempt to cover all the realistic regions of the parameter space.

$$\lambda \sim U(0,1), \quad \mu \sim U(0,\lambda), \quad f \sim U(0,1)$$

Note that under these conditions λ is always greater than μ.

We simulated datasets with 20, 40, 60, 80 and 100 taxa. For each number of taxa (*m*), we simulated 3960 topologies with random speciation rate (λ), random extinction rate (μ) and random fraction of sampled taxa (*f*). For each of these topologies, we then randomly choose a distribution from set *T* and sampled the branch lengths from this distribution (2*m*–2 branch lengths in total).

Other Python libraries that we used for the simulations are NumPy (Walt, et al. 2011), pandas (McKinney 2010) and ETE3 (Huerta-Cepas, et al. 2016). The python scripts for all simulations can be found on Github (https://github.com/suhanaser/empiricalGTRdist).

Inheritance Evolution: Inheritance Scheme Simulations

An evolutionary scenario would, ideally, have each lineage inheriting the parameters of its molecular evolutionary process from its parent lineage. At one extreme – where inheritance is perfect and the original evolutionary process is SRH, such a process would define a molecular evolutionary process that is SRH across the entire topology by simply defining a single SRH model at the root node. At the other extreme, where the association between parent and offspring lineages is no better than random and the original process is not SRH, there is no association between parent and offspring lineages and the process is maximally non-SRH. To mimic this situation, we designed a simulation approach that allows us to vary the homogeneity and stationarity assumptions both independently and together.

Our inheritance scheme allows us to vary the degree to which a single alignment has evolved under SRH conditions by simply adjusting the strength of inheritance of the substitution process and the base frequencies either jointly via a parameter we call ρ, or independently via parameters ν and ω respectively. When the inheritance parameters are set to 1 and the model at the root of the tree is reversible, the model will conform to SRH conditions. We can simulate increasing violation of SRH conditions simply by decreasing the inheritance parameters towards zero. When the relevant inheritance parameter is less than one, each branch inherits some proportion of its substitution model from the parent branch, while the remaining proportion of the model is selected at random from the empirical parameter distributions. In practice, the parameter in a descendant branch is calculated as the weighted sum of the parameter in the parent branch (where the weight is the inheritance parameter) and a randomly-generated parameter from the appropriate empirical distribution (where the weight is one minus the inheritance parameter).

We simulated data under five different categories of conditions using this scheme, in order to examine independently and together the effects of relaxing the stationarity and homogeneity assumptions.

1) SRH conditions (Fig. 1a).—In the simplest case for a model that conforms to the SRH assumptions, where model parameters are generated from the empirical distributions. This describes a model in which all branches inherit this reversible model from their parent branch without variation, such that all branches on the tree have the same reversible substitution model, conforming to the SRH assumptions.
2) Relaxing the stationarity assumption (Fig. 1b).—In order to hold the homogeneity assumption but relax the stationarity assumption, we introduce a parameter called ν $(0 \leq \nu \leq 1)$ that allows to vary the state frequency at the root while still keeping the same rate matrix for all branches of the tree. Mathematically, this can be described as:

$$\begin{cases} Q_i = Q_0 = \pi_0 S_0 & i \in \{\Phi\} \\ \pi_{root} = \nu^{d_{root}}\pi_0 + (1 - \nu^{d_{root}})\pi & \{\nu \in \mathbb{R}: 0 \leq \nu < 1\} \end{cases}$$

Where Q_i is the substitution rate matrix operating on branch i, and d_{root} is the branch length of the root branch.

When ν = 1, π_{root} is equal to π_0 and this scheme boils down to the first SRH condition. When ν = 0, π_{root} is equal to π, meaning that the root frequency is generated separately from π_0. π_{root} will vary between these two extremes when ν is between 0 and 1, with lower ν reflecting a larger deviation from stationary conditions.

3) Relaxing the homogeneity assumption (Fig. 1c).—In order to hold the stationarity assumption but relax the homogeneity assumption we need to simulate data in which ν is set to 1 (such that all branches have the same base frequencies as the root node), but we introduce a parameter ω that varies between zero and one (such that the inheritance of the parameters of the Q matrix ranges from completely random to near-perfect). We can describe this mathematically as follows:

$$\begin{cases} Q_i = \omega^{d_i}\pi_0 S_j + (1-\omega^{d_i})\pi_0 S & i,j \in \{\Phi\},\ \{\omega \in \mathbb{R}: 0 \le \omega < 1\} \\ \pi_{root} = \pi_0 \end{cases}$$

Where Q_i is the process operating on branch i, S_j are the substitution rates on the parent branch of branch i, and d_i is the branch length of the branch i.

4) Relaxing the stationarity and homogeneity assumptions simultaneously but independently (Fig. 1d).—We can simulate non-stationary and non-homogeneous data by setting both ν and ω to values less than one. When we relax both assumptions, we will allow Q_i and π_{root} to vary simultaneously but independently:

$$\begin{cases} Q_i = \omega^{d_i}\pi_0 S_j + (1-\omega^{d_i})\pi_0 S & i \in \{\Phi\},\ \{\omega \in \mathbb{R}: 0 \le \omega < 1\} \\ \pi_{root} = \nu^{d_{root}}\pi_0 + (1-\nu^{d_{root}})\pi & \{\nu \in \mathbb{R}: 0 \le \nu < 1\} \end{cases}$$

5) Relaxing the stationarity and homogeneity assumptions jointly (Fig. 1e).—While the 4th set of simulation conditions, above, allows us to vary homogeneity and stationarity jointly but independently, it suffers from the limitation that we have a maximum of two base frequency vectors in the tree ($\pi_{root}$and π_0). To relax this assumption further, we

will allow Q_i to vary while π_{root} stays fixed. In those settings, both homogeneity and stationarity will increase with ρ.

$$\begin{cases} Q_i = \rho\pi_0 S_0 + (1-\rho)\pi S & i \in \{\Phi\}, \ \{\rho \in \mathbb{R}: 0 \leq \rho \leq 1\} \\ \pi_{root} = \pi_0 \end{cases}$$

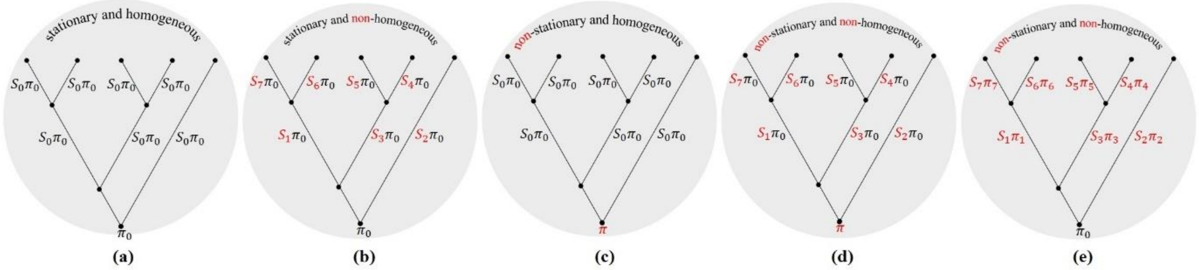


FIGURE 1. An example of 5 taxon tree with different degrees for homogeneity and stationarity. (a) stationary and homogeneous, **(b)** stationarity but not homogeneous, **(c)** non-stationary but homogeneous, **(d)** non-stationary and non-homogeneous where the stationarity and homogeneity assumptions are relaxed simultaneously but independently, **(e)** non-stationary and non-homogeneous where the stationarity and homogeneity assumptions are relaxed jointly.

Convergent Evolution: The Two-Matrix Scheme Simulations

Previous studies for simulating non-SRH evolution on phylogenies have used an approach in which two distantly related branches undergo severe but correlated changes in the molecular evolutionary process. To compare this approach to the more evolutionarily-motivated approach described above, we randomly chose two nodes that are not sisters and not descendants of each other and assigned a different rate matrix (denoted by $S_1\pi_1$) from the rest of the tree to all their descendant branches (Fig. 2).

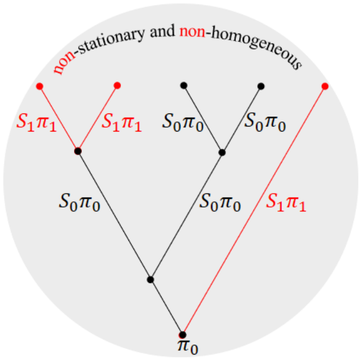


FIGURE 2. Non-stationary and non-homogeneous process with two different rate matrices $\mathbf{Q_0}$ and $\mathbf{Q_1}$.

Simulation Parameters

The simulation parameters that we use in this study are the strength of inheritance of the substitution process (ω), strength of inheritance of the base frequencies (ν), strength of inheritance of the substitution process and base frequencies (ρ), number of sites (n), and number of taxa (m) where the parameter space is:

$\omega, \nu, \rho \in \{0, 10^{-25}, 10^{-20}, 10^{-15}, 10^{-10}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1.0\}$

$n \in \{100, 1000, 10000\}$

$m \in \{20, 40, 60, 80, 100\}$

The inheritance weight parameters (ω, ν, ρ) were chosen to represent an even spread of corrected inheritance weights (i.e., the inheritance weights raised to the power of d, where d is the branch length) between zero and one. The number of taxa and number of sites are chosen to reflect the typical sizes of empirical datasets. For simulation under the inheritance scheme, we simulated 10 alignments of each combination of n, m, ν, and ω or n, m, and ρ for a total of 19,800 simulations. For simulation under the two-matrix scheme, we simulated 1000 alignments of each combination of n and m for a total of 15,000 simulations.

Tree Inference

Our first goal is to understand how the incorrect use of SRH models on data that have evolved under non-SRH processes can affect phylogenetic inference. To do this, we compare the tree topologies and branch lengths estimated with SRH models in IQ-TREE to the topologies and branch lengths used to simulate each dataset. For each simulated alignment, we ran IQ-TREE with ModelFinder (Kalyaanamoorthy, et al. 2017) and 1000 ultrafast bootstrap replicates (Hoang, et al. 2018). In order to assess the ability of SRH models to infer the correct tree topology we then compared the simulated tree topology to the estimated tree topology using three different metrics – normalized Robinson-Foulds distance (Robinson and Foulds 1981), Quartet distance (Estabrook, et al. 1985), and the Path-difference distance (Steel and Penny 1993). The normalized Robinson-Foulds distance between two trees is the fraction of internal branches that appear in one tree but not the other. It ranges from 0 to 1, where 0 means that the two trees are topologically identical and 1 means that the two trees have no branches in common. In order to assess the accuracy of branch length estimates, we tested whether the estimated branch lengths and the original branch lengths are drawn from the same distribution using the two-sample Kolmogorov-Smirnov test.

Detecting non-SRH Processes

We, therefore, tested the ability of three tests implemented in IQ-TREE to detect violation of the SRH assumptions: the MaxSymTests (Naser-Khdour, et al. 2019), the compositional Chi-square test, and the WvH test (Weiss and von Haeseler 2003). These three tests only need the composition of the alignment and therefore can be used with any analysis in IQ-TREE by adding the appropriate options to the command line, except for the Chi-square test that runs automatically for each alignment (Table 1).

Table 1. IQ-TREE option for each test

test	IQ-TREE option
MaxSym	--symtest
WvH	-m WHTEST
Chi-square	No option is needed

Since the Chi-square test tells us whether each sequence in the alignment fails the compositional homogeneity assumption, we use two different approaches that leverage the results of the Chi-square test (see also the Introduction):

1) A very conservative approach that we denote as Chi^2_{cons}. In this approach, we consider the alignment to fail the Chi-square test if one or more of the sequences in the alignment fails the test.
2) A less conservative ranking approach that we denote as Chi^2_{rank}. We record for each alignment what proportion of sequences that fail Chi-square test.

In the first case, we ask whether the proportion of replicate simulated alignments with one or more sequences failing the Chi-square test increases with the degree of violation of SRH conditions in the simulations. In the second case, we ask whether the proportion of sequences that fail the Chi-square test increases with the degree of violation of the SRH conditions in the simulations.

Results

Empirical Distributions

We derived the empirical distributions of the substitution model parameters, the nucleotide frequencies, and the proportion of invariant sites from 32,666 nucleotide alignments (Appendix

Table A.2). The empirical distribution of branch lengths we derived from 67 nucleotide and amino acid alignments consist of 33,178 partitions (Appendix Table A.1).

Using Kolmogorov-Smirnov test, we found the best-fit probability distribution for each one of these empirical distributions (Table 2, Appendix Table A.2, Appendix Figs. A.1-3).

TABLE 2. The best-fit probability distribution by Kolmogorov-Smirnov test

Parameter	Best-fit distribution	Shape (α)	Scale (β)	Location (x_0)
A↔C	Log-Laplace	1.695	1.636	-0.152
A↔G	Log-Laplace	1.465	4.930	-0.140
A↔T	Inverse-Weibull	3.015	1.841	-1.154
C↔G	Inverse-Weibull	1.793	1.651	-0.741
C↔T	Log-Laplace	1.551	5.182	-0.175
π_A	Generalized-logistic	0.557	0.026	0.313
π_T	Exponential-Weibull	0.843, 4.872	0.294	-0.001
π_C	Exponential-normal	1.769	0.027	0.173
π_G	Power-log-normal	0.090, 0.039	0.614	-0.471
%I	Beta	0.577, 4.707	2.162	-5.437
Branch length	Power-log-normal	1.208, 1.443	0.017	-7.1 e-05

Phylogenetic Inference is Unaffected by Violation of SRH Conditions in an inheritance Framework

Surprisingly, our results for the inheritance simulation scheme show that there is no detectable relationship between the severity with which SRH conditions were violated during the simulations and the accuracy of the tree topology or the tree length inferred from the simulated data. Specifically, we saw no relationship between the inheritance weight and the normalized RF (Robinson-Foulds), QD (Quartet Distance), or NPD (Normalized Path Difference) metrics in any of our inheritance simulations (Fig. 3, Appendix Figs. A.4-7). These metrics measure the difference between the inferred tree and the tree from which the alignment was simulated. If stronger violation of the SRH conditions affects phylogenetic inference we should expect to see that the distances are higher when the inheritance weight is lower, because a lower inheritance weight implies stronger model violation through less homogeneity (for the rate matrix) and less stationarity (for the base frequencies). In addition, our results show that the proportion of simulated datasets for which the simulated tree is recovered from the simulated alignment is constant at around 0.25 in the inheritance scheme simulations regardless of the inheritance weight (Fig. 3, Appendix Fig. A.4). Finally, we see no correlation between the inheritance weights and the proportion of datasets that fail a Kolmogorov-Smirnov test comparing the true and estimated branch lengths, suggesting that violation of SRH assumptions in our evolutionary framework has no detectable effect on the estimation of branch lengths (Fig. 4, Appendix Fig. A.19).

Tree Topologies, but not Branch Lengths, are Affected by Severe and Convergent Violation of SRH Conditions

Our results show that convergent violation of SRH assumptions by allowing two distantly related branches to have identical substitution models has increasingly severe effects on phylogenetic inference as the severity of the changes in the substitution models increases. Under the two-matrix scheme, we expect to see higher distances between the true tree and the estimated tree when there are larger Euclidian distances between the original matrix and the matrix under which the divergent clades evolve. In two out of the three metrics (Robison-Foulds and Path-Difference) we found a weak but significant correlation between the distance between the matrices and the distance between the topologies (Fig. 3, Appendix Fig. A.4, Appendix Figs. A.8-10). However, in the third metric (Quartet Distance) we found no correlation. Notably, the distance between the true tree and the estimated tree increases only when the Euclidean distance between the two matrices is very high. Nevertheless, the proportion of simulated datasets for which the simulated tree is recovered from the simulated alignment declines exponentially as the difference between the matrices in the two-matrix scheme increases (Fig. 3).

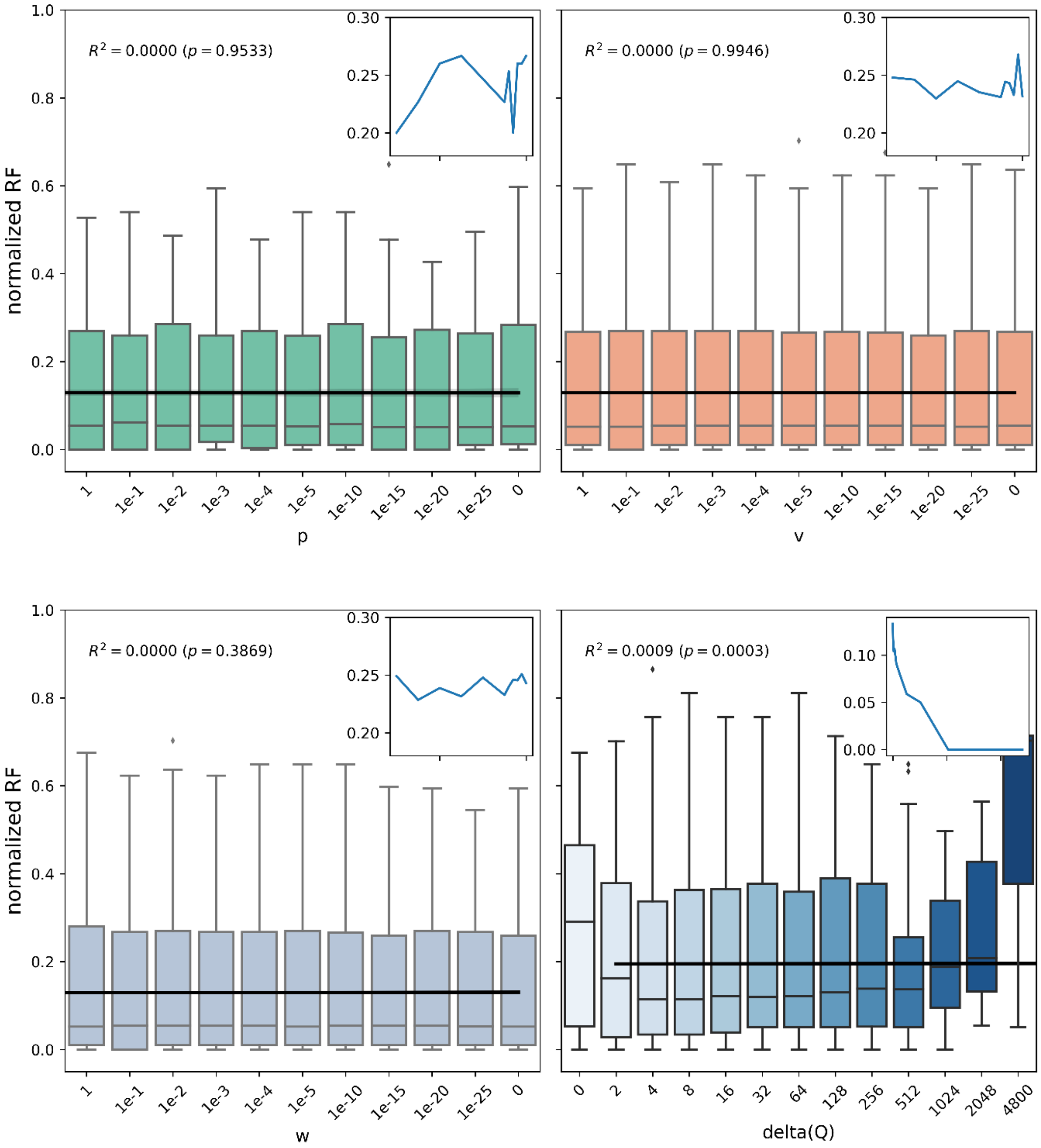


FIGURE 3. Normalized Robinson-Foulds distance between the estimated tree topology and the original tree topology as a function of the inheritance weight (ν, ω ,ρ) in the first simulation scheme, and the distance between the two matrices (delta(Q)) in the second simulation scheme. The small plots show the proportion of datasets in which the distance between the estimated topology and the original topology equals zero as a function of the inheritance weight and the distance between the two matrices. If violation of SRH model assumptions increases topological error, we expect the nRF distance to increase towards the right of each plot. The figure shows that for the first simulation scheme, which mimics a stochastic evolutionary process, there is no detectable association between violation of SRH conditions and topological

error. For the second simulation scheme, which mimics an extreme convergent situation, topological error increases with increasing violation of SRH conditions.

The proportion of analyses in which the simulated tree is recovered positively declines from around 0.20 when there is no model violation to zero when the Euclidean distance between the matrices is around 2000, confirming that even the lowest levels of SRH violation have detectable negative effects on phylogenetic inference under the two-matrix scheme.

Finally, we see no correlation between the Euclidean distance between the two matrices and the proportion of datasets that fail a Kolmogorov-Smirnov test comparing the distributions of the true and estimated branch lengths, suggesting that violation of SRH assumptions in the convergent framework has limited effects on the estimation of branch lengths (Fig. 4, Appendix Fig. A.22).

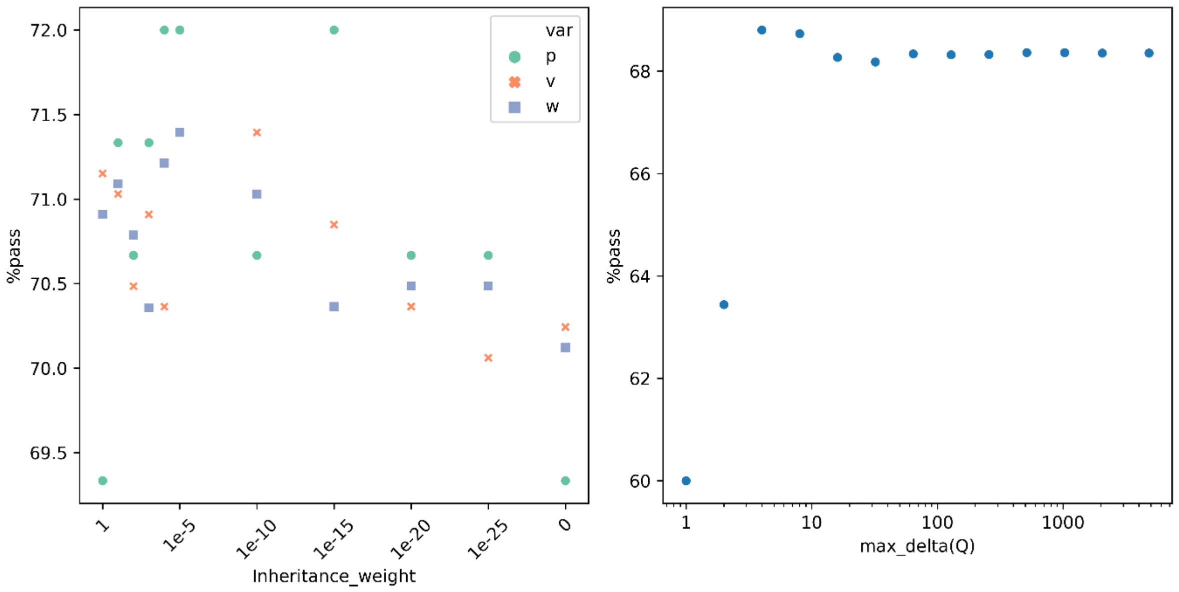


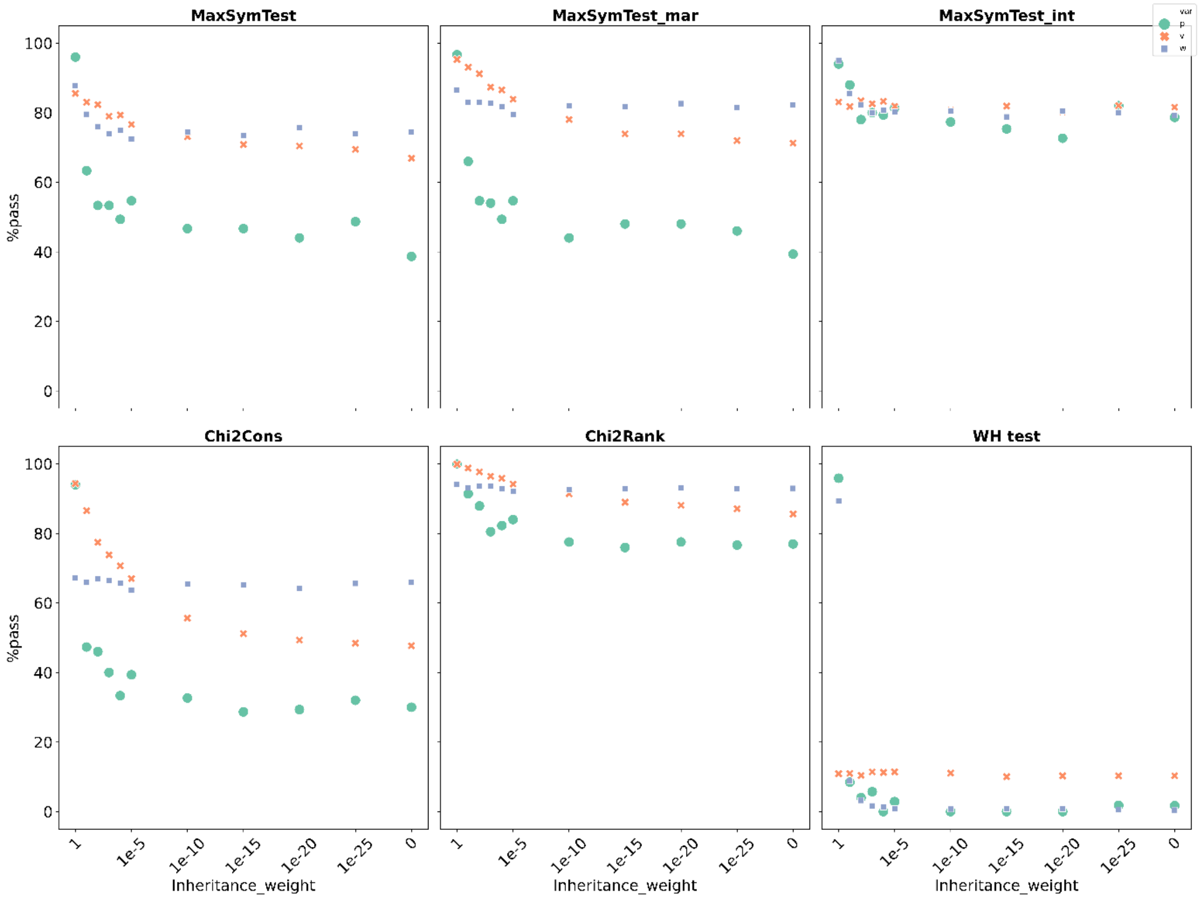
FIGURE 4. The percentage of datasets that pass the KS test (a dataset passes when there is no evidence to suggest that the inferred branch length distributions differ) as a function of the inheritance weights (ν,ρ,ω) (left-hand panel) and maximum Euclidian distance between the two matrices used to simulate the data (right-hand panel).

Tests for Detecting non-SRH Processes are Successful but have High False-Negative Rates

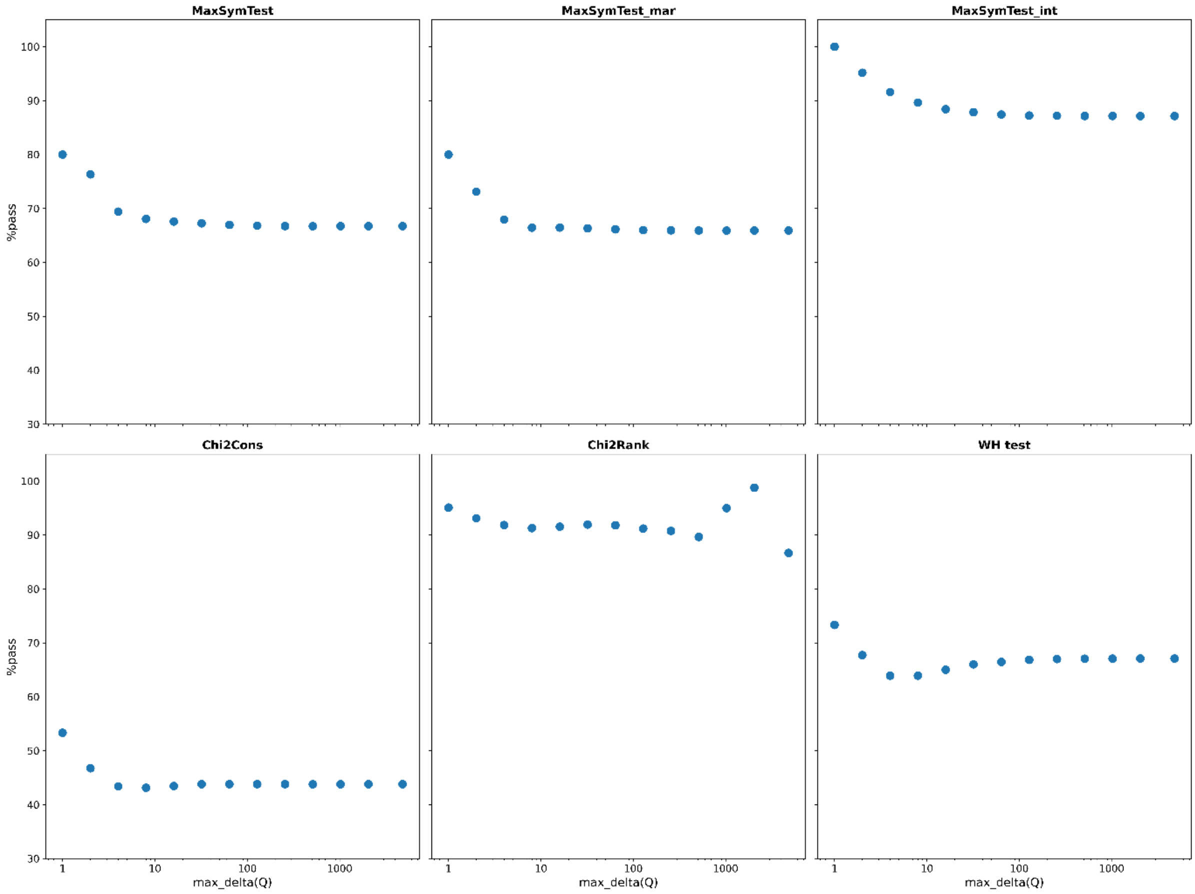
As expected, the ability of all three MaxSym tests to reject the null hypothesis of stationarity and homogeneity improves as the inheritance weight in the evolutionary simulations decreases (i.e. as the violation of SRH conditions increases), the distance between the two matrices in the convergent simulations increases, and the number of sites in the alignment increases (Fig. 5, Appendix Figs. A.11-13). Moreover, the three MaxSym tests have a reasonable false positive rate of approximately 4.5% (Appendix Table A.4). However, they also have very high false-negative rates of 50-90%, depending on the test and the particular simulation conditions (Fig. 5, Appendix Table A.3, Appendix Figs. A.14-16). In the two-matrix scheme simulations, the false-negative rates of MaxSym, MaxSym$_{\mathrm{mar}}$, and MaxSym$_{\mathrm{int}}$ tests are 67%, 66%, and 87%, respectively. Thus, across all simulation conditions, a significant result from a MaxSymTest can be reliably interpreted as indicating that an alignment violates the SRH conditions, but the test will fail to identify many such alignments.

Similarly to the MaxSym tests results, the Chi_{cons}^{2} and Chi_{Rank}^{2} tests show an increase in the proportion of alignments and/or sequences that fail the test in each dataset as the inheritance weight decreases, and the number of sites increases (Fig. 5a, Appendix Fig. A.17). The false-positive rates of the Chi_{cons}^{2} test is 6% (Appendix Table A.6). The false-negative rate of the Chi_{cons}^{2} test in the inheritance-scheme simulations is 57% (Appendix Table A.5). Moreover, similar to the MaxSym tests, in the two-matrix scheme simulation, the percentage of datasets that pass the Chi_{cons}^{2} decreases logarithmically the higher the distance between the two matrices (Fig. 5b, Appendix Fig. A.18). The false-negative rate of the Chi_{cons}^{2} test under extreme convergent evolution is the smallest of all the tests considered here under these conditions, and it is around 44% (Appendix Table A.5).

In the inheritance-scheme simulations, similarly to the MaxSym tests, the Chi^2_{cons} and Chi^2_{Rank}, the WvH test shows an increase in the proportion of alignments that fail the test as the inheritance weight (ω and ρ) decreases. However, ν has no effect on the proportion of alignments that fail the WvH test. The false-positive rate of the WvH test is 3.5% (Appendix Table A.8), which is lower than any of the MaxSym tests or the Chi^2_{cons} test. In addition, the false-negative rate of the WvH test (Appendix Table A.7) in the inheritance-scheme simulations is lower than all the other tests (~30%) but it is still high under the two-matrix scheme simulations (~67%). Yet, due to numerical instability, the WvH test could be only applied to half of the datasets in the two simulation schemes.



(a)



(b)

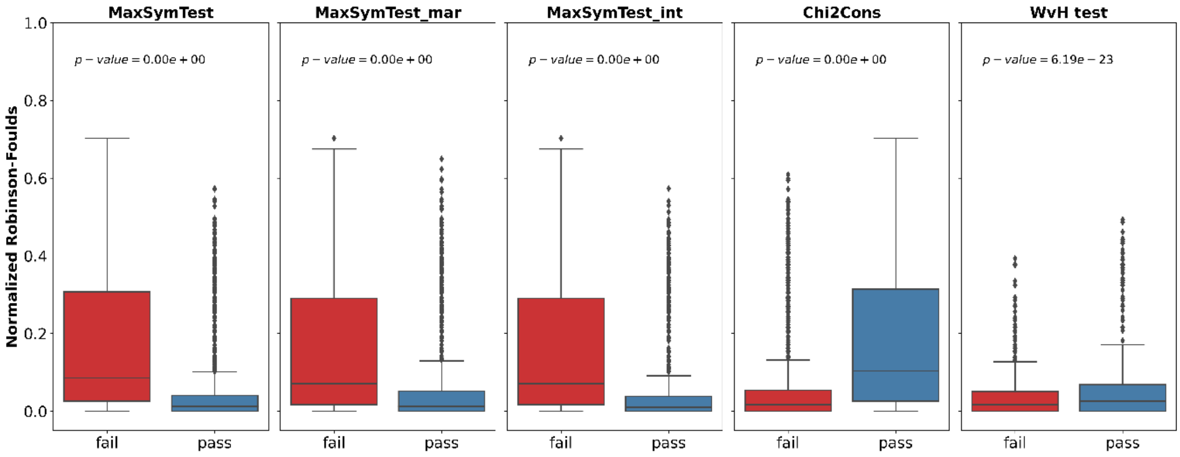
FIGURE 5. The mean percentage of datasets that pass each of the MaxSym tests, the WvH test, and the Chi-square test as a function of **(a)** the inheritance weights (ν,ρ,ω) **(b)** maximum Euclidian distance between the two matrices. We define datasets that pass the Chi2Cons test as datasets where all the sequences pass the Chi-square test. On the other hand, Chi2Rank shows the proportion of sequences that pass the Chi-square test in each dataset.

MaxSymTest$_{\text{int}}$ is a good predictor of correct tree inference

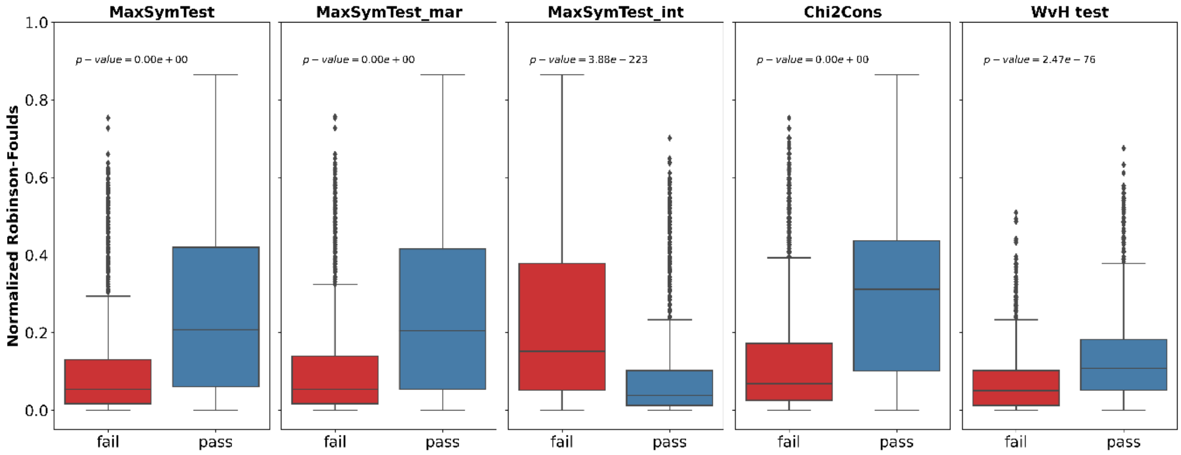
A key question for empiricists is whether tests of model adequacy are likely to improve phylogenetic inference. To explore this in our simulation framework, we asked whether datasets that are rejected by the tests we evaluated tended to be associated with more phylogenetic tree error than those that were not rejected. To do this, we used three different metrics of tree distance (the normalized Robison-Foulds (RF), Path-Difference, and Quartet distance) and asked whether datasets that fail the test (i.e. have detectable non-SRH processes) tended to result in trees that were further from the true tree (i.e. had higher nRF distances) when analysed using SRH models. All three showed very similar results, so we show the normalized Robinson-Foulds results here (Fig. 6) and the other metrics in the supplementary information (Appendix Fig. A.23a, Appendix Fig. A.24a).

For the inheritance scheme simulations, we found as expected that datasets that failed the MaxSym tests were associated with trees much further from the true tree than those that passed the tests, although there was substantial variation within each category (Fig. 6a, Appendix Fig. A.23a, Appendix Fig. A.24a). Surprisingly, this pattern was reversed for the Chi^2_{cons} test, and there was a very small difference in tree distances with the WvH test (Fig. 6a). Welch's t-test results suggest all of the differences are statistically significant (p<<0.05, Fig. 6a).

For the two-matrix simulations, the only test for which datasets that failed were associated with trees further from the true tree was the MaxSym$_{int}$ test (Fig. 6b, Appendix Fig. A.23b, Appendix Fig. A.24b). For all other tests, datasets that failed the test were associated with trees that were markedly *closer* to the true tree than datasets that passed the tests (Fig. 6b, Appendix Fig. A.23b, Appendix Fig. A.24b). Again, Welch's t-test results suggest all of the differences are statistically significant (p<<0.05, Fig. 6b).



(a)



(b)

FIGURE 6. The normalized Robison-Foulds distance as a function of passing/failing the tests in **(a)** the inheritance scheme simulations and **(b)** the two-matrix scheme simulations. The p-value in each sub-figure is calculated from the Welch's t-test statistics.

Discussion

Using two different simulation schemes, we explored the impact of violating the assumption of evolution under stationary, reversible, and homogeneous (SRH) conditions on ML phylogenetic tree inference. Our study extends the simulations in many previous studies by simulating data under an evolutionary scenario in which molecular evolutionary models evolve along a phylogeny. Our results show that the inference of phylogenetic tree topologies and branch lengths are surprisingly robust to violations of SRH assumptions under an evolutionary scheme. But similarly to previous studies, we show that in extreme cases of convergent molecular evolution the incorrect assumption of SRH conditions can severely mislead phylogenetic inference.

The first simulation scheme we introduced in this paper, which we called *the inheritance scheme*, allows tree branches to inherit their substitution process from their ancestor. The second simulation scheme, which we called *the two-matrix scheme*, is similar to previous studies and allows two distantly related monophyletic sub-trees to evolve with a different evolutionary process from the rest of the tree (Galtier and Gouy 1995; Jermiin, et al. 2004; Jayaswal, et al. 2011; Duchene, et al. 2017).

Surprisingly, our results show no correlation between errors in the topology or branch length inference and any of the inheritance scheme parameters, even in extreme cases where the evolutionary process is completely heterogeneous and non-stationary. These results indicate that ML tree inference with SRH models is surprisingly robust to even quite extreme violations of the SRH conditions.

Under the two-matrix simulation scheme, we found a small but significant increase in topological inference error and the extent of the violation of the SRH assumptions. Specifically, the more extreme the evolutionary convergence, the larger the errors in the topological inference that assumes SRH conditions. Despite this, we found no correlation between branch length inference and the distance between the two matrices. These results emphasize the limitations of ML inference to operate under certain model violations, especially when these violations are highly imbalanced along the tree, as in the case of the two-matrix scheme simulations. These results indicate that the inference of the substitution model is more influenced by the imbalance of the model violation distribution along the tree than by the model violation itself. This conclusion agrees well with all previous simulation studies of similar simulation conditions (e.g. Jermiin, et al. 2004; Duchene, et al. 2017; Jermiin, et al. 2019).

In this study, we also tested the power of the MaxSym tests, WvH test and two variations of the Chi-squared test to detect model violation due to non-SRH evolution. Our results show that those tests were able to detect some model violations in both simulation schemes. As expected, the power of all tests to detect model violation due to non-SRH evolution improves dramatically as alignment length increases, reflecting simply the larger amount of information available in longer alignments. However, the power of most of the tests we looked at was somewhat limited – even in the best-case scenario when violation of the SRH conditions was severe, most tests were able to detect this violation in less than 50% of the simulated datasets (Fig. 5). The two exceptions were the WvH test, which was able to detect the vast majority of datasets simulated with model violation under the inheritance scheme simulations (Fig. 6a) and the conservative Chi-Square test, which was able to detect the majority of datasets simulated with model violation under the convergent evolution scheme. However, the WvH test could not be applied to half of the datasets in our simulations due to numerical instability, suggesting that it may be less useful for detecting violations of SRH conditions in practice than the other tests.

The high false-negative rates of the MaxSym tests also suggest that some of the partitions that violate the SRH assumption are not detected by those tests which means that the impact of the model violation on the phylogenetic inference is actually higher than it seems. The implications are big also for the empirical datasets from Chapter 1; if the tests were more powerful I would expect to see more extreme results.

The utility of any test of model adequacy in practice is likely to be tied to the amount of phylogenetic error that a test helps empiricists avoid. All models used in phylogenetic analyses are gross oversimplifications of highly complex molecular evolutionary processes, and so merely detecting violations of models is necessary but not sufficient for a model adequacy test to be useful. Because of this, we asked for each test whether the datasets that fail the test were associated with more or less topological inference error than the datasets that passed the test. Surprisingly, the only test that performed consistently well in this regard was the MaxSymTest$_{int}$. Under the inheritance scheme simulations, all three MaxSym tests are good predictors of phylogenetic accuracy; trees that pass any of those tests are closer to the true tree than trees that fail. The WvH and Chi^2_{cons} tests on the other hand are bad predictors of phylogenetic accuracy; trees that *fail* the Chi^2_{cons} test are usually closer to the true tree, while there is only a small difference between trees that fail and trees that pass the WvH test. Surprisingly, under the convergent simulation scheme, the MaxSymTest$_{\text{int}}$ is the *only* test for which datasets that pass the test are closer to the true tree than datasets that fail the test (Fig. 6). For all other tests, the datasets that pass the test were substantially *further* from the true tree than those that fail the test.

It is challenging to disentangle why some tests of the SRH assumptions tend to detect datasets that are associated with more topological error, while others show the opposite tendency (Fig. 6), although we suspect this is often driven by the interplay of the power of the tests, phylogenetic signal, and stochastic error in tree estimation. Across all simulation conditions, the only test which consistently showed the desirable behaviour from an empirical standpoint (i.e. where datasets that fail the test are associated with more topological error) was the MaxSymTest$_{\text{int}}$. All other tests showed evidence of having the opposite tendency (Fig. 6) in at least some simulation conditions. In the case of the WvH test, for which alignments that fail the test were consistently associated with *less* topological inference error when analysed under SRH models, we suspect that the underlying reason may be the reliance of the test on a parametric bootstrap. The WvH test depends fundamentally on a tree estimated with an SRH model to estimate the null distribution of the test statistic. If this tree is wrong, as we show occurs under model violation, then the null distribution may be incorrect and the test misled. For the other tests we suspect that the tendency is driven largely by the fact that datasets with few informative sites will tend to both pass the tests *and* be associated with high topological error, with both caused by the limited information in the data, although further work is needed to understand these relationships in more detail. Nevertheless, the observation that across all simulation conditions, datasets that fail the MaxSymTest$_{\text{int}}$ are associated with higher topological error do suggest that violations of the *homogeneity* assumption might be the most important when it comes to phylogenetic inference with SRH models, since the MaxSymTest$_{int}$ tests primarily for violations of homogeneity.

These results combined with the results from the inheritance scheme simulations, emphasize the need to use different methods and tests for model violation in phylogenetic analyses since each test can capture a different aspect of model violation. A ~~new~~ phylogenetic protocol (Jermiin, et al. 2020) stresses the need to validate the assumptions of the models in advance. If the data in hand violates the model's assumptions then different models or methods should be considered. A surprising result from this work is that the MaxSymTest$_{int}$ is a good predictor for phylogenetic accuracy. Yet, one should bear in mind that this test has the highest false-negative rate among all of the tests examined in this study.

It is noteworthy that our results from the different simulation schemes agree with the results from empirical data (Naser-Khdour, et al. 2019). They emphasize the impact of model violation due to non-SRH evolution on phylogenetic inference and suggest that reducing model violation in phylogenetic analysis by using the protocol of phylogenetic inference (Jermiin, et al. 2020) or using more complex substitution models e.g.(Galtier and Gouy 1998; Tamura and Kumar 2002; Blanquart and Lartillot 2008; Dutheil, et al. 2012; Zou, et al. 2012; Groussin, et al. 2013; Jayaswal, et al. 2014) has the potential to improve phylogenetic accuracy.

For the purpose of this study, in order to simulate data that mimic as closely as possible empirical alignments, we extracted the empirical distributions of base frequencies, substitution rates, proportion of invariable sites, and branch lengths from tens of thousands of empirical datasets. In addition to their use in this paper, these empirical distributions, along with their best-fit distributions may be useful for a wide variety of simulation studies, or for specifying prior distributions for Bayesian phylogenetic methods.

Funding

This work was supported by an Australian Research Council (Grant No. DP200103151 to R.L., B.Q.M.) and by a Chan-Zuckerberg Initiative grant to B.Q.M and R.L.

References

Ababneh F, Jermiin LS, Ma C, Robinson J. 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics 22:1225-1231.

Anderson FE, Bergman A, Cheng SH, Pankey MS, Valinassab T. 2013. Data from: Lights out: the evolution of bacterial bioluminescence in Loliginidae. In: Dryad Data Repository.

Anderson FE, Bergman A, Cheng SH, Pankey MS, Valinassab T. 2014. Lights out: the evolution of bacterial bioluminescence in Loliginidae. Hydrobiologia 725:189-203.

Aouad M, Taib N, Oudart A, Lecocq M, Gouy M, Brochier-Armanet C. 2018. Extreme halophilic archaea derive from two distinct methanogen Class II lineages. Mol Phylogenet Evol 127:46-54.

Ballesteros JA, Sharma PP. 2019a. A Critical Appraisal of the Placement of Xiphosura (Chelicerata) with Account of Known Sources of Phylogenetic Error. Syst. Biol.

Ballesteros JA, Sharma PP. 2019b. Data from: A critical appraisal of the placement of Xiphosura (Chelicerata) with account of known sources of phylogenetic error. In: Dryad.

Barry D, Hartigan JA. 1987. Statistical Analysis of Hominoid Molecular Evolution. Statistical Science 2:191-207.

Becker EA, Yao AI, Seitzer PM, Kind T, Wang T, Eigenheer R, Shao KS, Yarov-Yarovoy V, Facciotti MT. 2016. A Large and Phylogenetically Diverse Class of Type 1 Opsins Lacking a Canonical Retinal Binding Site. PLoS One 11:e0156543.

Becker EA, Yao AI, Seitzer PM, Kind T, Wang T, Eigenheer R, Shao KSY, Yarov-Yarovoy V, Facciotti MT. 2017. Data from: A large and phylogenetically diverse class of type 1 opsins lacking a canonical retinal binding site. In: Dryad.

Bergsten J, Nilsson AN, Ronquist F. 2013a. Bayesian tests of topology hypotheses with an example from diving beetles. Syst. Biol. 62:660-673.

Bergsten J, Nilsson AN, Ronquist F. 2013b. Data from: Bayesian tests of topology hypotheses with an example from diving beetles. In: Dryad Data Repository.

Betancur RR, Li C, Munroe TA, Ballesteros JA, Orti G. 2013. Addressing gene tree discordance and non-stationarity to resolve a multi-locus phylogeny of the flatfishes (Teleostei: Pleuronectiformes). Syst. Biol. 62:763-785.

Blanquart S, Lartillot N. 2008. A site- and time-heterogeneous model of amino acid replacement. Mol. Biol. Evol. 25:842-858.

Borowiec ML, Lee EK, Chiu JC, Plachetzki DC. 2016. Data from: Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. In: Dryad Digital Repository.

Borowiec ML, Lee EK, Chiu JC, Plachetzki DC. 2015. Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. BMC Genomics 16:987.

Boussau B, Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. Syst. Biol. 55:756-768.

Branstetter MG, Danforth BN, Pitts JP, Faircloth BC, Ward PS, Buffington ML, Gates MW, Kula RR, Brady SG. 2017a. Data from: Phylogenomic insights into the evolution of stinging wasps and the origins of ants and bees. In: Dryad Digital Repository.

Branstetter MG, Danforth BN, Pitts JP, Faircloth BC, Ward PS, Buffington ML, Gates MW, Kula RR, Brady SG. 2017b. Phylogenomic Insights into the Evolution of Stinging Wasps and the Origins of Ants and Bees. Curr. Biol. 27:1019-1025.

Broughton RE, Betancur RR, Li C, Arratia G, Orti G. 2013a. Data from: Multi-locus phylogenetic analysis reveals the pattern and tempo of bony fish evolution. In: Dryad Data Repository.

Broughton RE, Betancur RR, Li C, Arratia G, Orti G. 2013b. Multi-locus phylogenetic analysis reveals the pattern and tempo of bony fish evolution. PLoS Curr 5.

Brown JM, ElDabaje R. 2009. PuMA: Bayesian analysis of partitioned (and unpartitioned) model adequacy. Bioinformatics 25:537-538.

Brown RM, Siler CD, Das I, Min PY. 2012a. Data from: Testing the phylogenetic affinities of Southeast Asia’s rarest geckos: Flap-legged geckos (Luperosaurus), Flying geckos (Ptychozoon) and their relationship to the pan-Asian genus Gekko. In: Dryad Data Repository.

Brown RM, Siler CD, Das I, Min Y. 2012b. Testing the phylogenetic affinities of Southeast Asia's rarest geckos: Flap-legged geckos (Luperosaurus), Flying geckos (Ptychozoon) and their relationship to the pan-Asian genus Gekko. Mol Phylogenet Evol 63:915-921.

Cannon JT, Vellutini BC, Smith J, 3rd, Ronquist F, Jondelius U, Hejnol A. 2016a. Xenacoelomorpha is the sister group to Nephrozoa. Nature 530:89-93.

Cannon JT, Vellutini BC, Smith J, Ronquist F, Jondelius U, Hejnol A. 2016b. Data from: Xenacoelomorpha is the sister group to Nephrozoa. In: Dryad Data Repository.

Chen M-Y, Liang D, Zhang P. 2015a. Data from: Selecting question-specific genes to reduce incongruence in phylogenomics: a case study of jawed vertebrate backbone phylogeny. In: Dryad.

Chen MY, Liang D, Zhang P. 2015b. Selecting Question-Specific Genes to Reduce Incongruence in Phylogenomics: A Case Study of Jawed Vertebrate Backbone Phylogeny. Syst. Biol. 64:1104-1120.

Chernomor O, von Haeseler A, Minh BQ. 2016. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. Syst. Biol. 65:997-1008.

Cognato AI, Vogler AP. 2001a. Data from: Exploring data interaction and nucleotide alignment in a multiple gene analysis of Ips (Coleoptera: Scolytinae). In: Dryad Data Repository.

Cognato AI, Vogler AP. 2001b. Exploring data interaction and nucleotide alignment in a multiple gene analysis of Ips (Coleoptera: Scolytinae). Syst. Biol. 50:758-780.

Collins TM, Fedrigo O, Naylor GJ. 2005. Choosing the best genes for the job: the case for stationary genes in genome-scale phylogenetics. Syst. Biol. 54:493-500.

Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC. 2012a. Data from: More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. In: Dryad Digital Repository.

Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC. 2012b. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biol. Lett. 8:783-786.

Day JJ, Peart CR, Brown KJ, Friel JP, Bills R, Moritz T. 2013a. Continental diversification of an African catfish radiation (Mochokidae: Synodontis). Syst. Biol. 62:351-365.

Day JJ, Peart CR, Brown KJ, Friel JP, Bills R, Moritz T. 2013b. Data from: Continental diversification of an African catfish radiation (Mochokidae: Synodontis). In: Dryad Data Repository.

Devitt TJ, Cameron Devitt SE, Hollingsworth BD, McGuire JA, Moritz C. 2013. Data from: Montane refugia predict population genetic structure in the Large-blotched Ensatina salamander. In: Dryad Data Repository.

Devitt TJ, Devitt SE, Hollingsworth BD, McGuire JA, Moritz C. 2013. Montane refugia predict population genetic structure in the Large-blotched Ensatina salamander. Mol. Ecol. 22:1650-1665.

Dornburg A, Moore JA, Webster R, Warren DL, Brandley MC, Iglesias TL, Wainwright PC, Near TJ. 2012a. Data from: Molecular phylogenetics of squirrelfishes and soldierfishes (Teleostei:Beryciformes: Holocentridae): reconciling more than 100 years of taxonomic confusion. In: Dryad Data Repository.

Dornburg A, Moore JA, Webster R, Warren DL, Brandley MC, Iglesias TL, Wainwright PC, Near TJ. 2012b. Molecular phylogenetics of squirrelfishes and soldierfishes (Teleostei: Beryciformes: Holocentridae): reconciling more than 100 years of taxonomic confusion. Mol Phylogenet Evol 65:727-738.

Duchene DA, Duchene S, Ho SYW. 2017. New Statistical Criteria Detect Phylogenetic Bias Caused by Compositional Heterogeneity. Mol. Biol. Evol. 34:1529-1534.

Dutheil JY, Galtier N, Romiguier J, Douzery EJ, Ranwez V, Boussau B. 2012. Efficient selection of branch-specific models of sequence evolution. Mol. Biol. Evol. 29:1861-1874.

Estabrook GF, McMorris F, Meacham CA. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. Systematic Zoology 34:193-200.

Faircloth BC, Sorenson L, Santini F, Alfaro ME. 2013a. Data from: A phylogenomic perspective on the radiation of ray-finned fishes based upon targeted sequencing of ultraconserved elements (UCEs). In: Dryad Data Repository.

Faircloth BC, Sorenson L, Santini F, Alfaro ME. 2013b. A Phylogenomic Perspective on the Radiation of Ray-Finned Fishes Based upon Targeted Sequencing of Ultraconserved Elements (UCEs). PLoS One 8:e65923.

Fan L, Wu D, Goremykin V, Xiao J, Xu Y, Garg S, Zhang C, Martin WF, Zhu R. 2020. Phylogenetic analyses with systematic taxon sampling show that mitochondria branch within Alphaproteobacteria. Nat Ecol Evol 4:1213-1219.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368-376.

Felsenstein J. 2004. Inferring Phylogenies. Sunderland, Massachusetts: Sinauer Associates, Inc.

Felsenstein J. 1983. Statistical inference of phylogenies. Journal of the Royal Statistical Society: Series A (General) 146:246-262.

Fong JJ, Brown JM, Fujita MK, Boussau B. 2012a. Data from: A phylogenomic approach to vertebrate phylogeny supports a turtle-archosaur affinity and a possible paraphyletic Lissamphibia. In: Dryad Data Repository.

Fong JJ, Brown JM, Fujita MK, Boussau B. 2012b. A phylogenomic approach to vertebrate phylogeny supports a turtle-archosaur affinity and a possible paraphyletic lissamphibia. PLoS One 7:e48990.

Foster PG. 2004. Modeling compositional heterogeneity. Syst. Biol. 53:485-495.

Foster PG, Hickey DA. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J. Mol. Evol. 48:284-290.

Galtier N, Gouy M. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. 15:871-879.

Galtier N, Gouy M. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proc Natl Acad Sci U S A 92:11317-11321.

Groussin M, Boussau B, Gouy M. 2013. A branch-heterogeneous model of protein evolution for efficient inference of ancestral sequences. Syst. Biol. 62:523-538.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160-174.

Hillis DM, Huelsenbeck JP, Cunningham CW. 1994. Application and accuracy of molecular phylogenies. Science 264:671-677.

Ho JW, Adams CE, Lew JB, Matthews TJ, Ng CC, Shahabi-Sirjani A, Tan LH, Zhao Y, Easteal S, Wilson SR, et al. 2006. SeqVis: visualization of compositional heterogeneity in large alignments of nucleotides. Bioinformatics 22:2162-2163.

Ho SY, Jermiin L. 2004. Tracing the decay of the historical signal in biological sequence data. Syst. Biol. 53:623-637.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. 35:518-522.

Horn JW, Xi Z, Riina R, Peirson JA, Yang Y, Dorsey BL, Berry PE, Davis CC, Wurdack KJ. 2014a. Data from: Evolutionary bursts in Euphorbia (Euphorbiaceae) are linked with photosynthetic pathway. In: Dryad Data Repository.

Horn JW, Xi Z, Riina R, Peirson JA, Yang Y, Dorsey BL, Berry PE, Davis CC, Wurdack KJ. 2014b. Evolutionary bursts in Euphorbia (Euphorbiaceae) are linked with photosynthetic pathway. Evolution 68:3485-3504.

Hua X, Lanfear R. 2018. The influence of non-random species sampling on macroevolutionary and macroecological inference from phylogenies. Methods in Ecology and Evolution 9:1353-1362.

Huelsenbeck JP, Hillis DM. 1993. Success of Phylogenetic Methods in the Four-Taxon Case. Syst. Biol. 42:247-264.

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol. Biol. Evol. 33:1635-1638.

Irisarri I, Baurain D, Brinkmann H, Delsuc F, Sire J, Kupfer A, Petersen J, Jarek M, Meyer A, Vences M, et al. 2017a. Data from: Phylotranscriptomic consolidation of the jawed vertebrate timetree. In: Dryad Digital Repository.

Irisarri I, Baurain D, Brinkmann H, Delsuc F, Sire JY, Kupfer A, Petersen J, Jarek M, Meyer A, Vences M, et al. 2017b. Phylotranscriptomic consolidation of the jawed vertebrate timetree. Nat Ecol Evol 1:1370-1378.

Jarvis ED, Mirarab S, Aberer A, Houde P, Li C, Ho S, Faircloth BC, Nabholz B, Howard JT, Suh A, et al. 2014. Data from: Phylogenomic analyses data of the avian phylogenomics project. In: GigaScience Database.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT, et al. 2015. Phylogenomic analyses data of the avian phylogenomics project. Gigascience 4:4.

Jayaswal V, Ababneh F, Jermiin LS, Robinson J. 2011. Reducing model complexity of the general Markov model of evolution. Mol. Biol. Evol. 28:3045-3059.

Jayaswal V, Wong TK, Robinson J, Poladian L, Jermiin LS. 2014. Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. Syst. Biol. 63:726-742.

Jermiin L, Ho SY, Ababneh F, Robinson J, Larkum AW. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst. Biol. 53:638-643.

Jermiin LS, Catullo RA, Holland BR. 2020. A new phylogenetic protocol: dealing with model misspecification and confirmation bias in molecular phylogenetics. NAR Genom Bioinform 2:lqaa041.

Jermiin LS, Ho JWK, Lau KW, Jayaswal V. 2009. SeqVis: a tool for detecting compositional heterogeneity among aligned nucleotide sequences. In. Bioinformatics for DNA sequence analysis: Springer. p. 65-91.

Jermiin LS, Lovell DR, Misof B, Foster PG, Robinson J. 2019. Software for Detecting Heterogeneous Evolutionary Processes across Aligned Sequence Data. bioRxiv:828996.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14:587-589.

Kawahara AY, Rubinoff D. 2013a. Convergent evolution of morphology and habitat use in the explosive Hawaiian fancy case caterpillar radiation. J. Evol. Biol. 26:1763-1773.

Kawahara AY, Rubinoff D. 2013b. Data from: Convergent evolution in the explosive Hawaiian Fancy Cased caterpillar radiation. In: Dryad Data Repository.

Kimura M. 1980. A Simple Method for Estimating Evolutionary Rates of Base Substitutions through Comparative Studies of Nucleotide-Sequences. J. Mol. Evol. 16:111-120.

Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K. 2012. Statistics and truth in phylogenomics. Mol. Biol. Evol. 29:457-472.

Kumar S, Gadagkar SR. 2001. Disparity index: a simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. Genetics 158:1321-1327.

Lanave C, Preparata G, Saccone C, Serio G. 1984. A new method for calculating evolutionary substitution rates. J. Mol. Evol. 20:86-93.

Lanave C, Tommasi S, Preparata G, Saccone C. 1986. Transition and transversion rate in the evolution of animal mitochondrial DNA. BioSyst. 19:273-283.

Lartillot N, Delsuc F. 2012a. Data from: Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. In: Dryad Data Repository.

Lartillot N, Delsuc F. 2012b. Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. Evolution 66:1773-1787.

Leache AD, Chavez AS, Jones LN, Grummer JA, Gottscho AD, Linkem CW. 2015. Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. Genome Biol Evol 7:706-719.

Leaché AD, Chavez AS, Jones LN, Grummer JA, Gottscho AD, Linkem CW. 2015. Data from: Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. In: Dryad.

Liu ZQ, Liu YF, Kuermanali N, Wang DF, Chen SJ, Guo HL, Zhao L, Wang JW, Han T, Wang YZ, et al. 2018. Sequencing of complete mitochondrial genomes confirms synonymization of Hyalomma asiaticum asiaticum and kozlovi, and advances phylogenetic hypotheses for the Ixodidae. PLoS One 13:e0197524.

Lockhart PJ, Steel MA, Hendy MD, Penny D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11:605-612.

Looney BP, Ryberg M, Hampe F, Sanchez-Garcia M, Matheny PB. 2016. Into and out of the tropics: global diversification patterns in a hyperdiverse clade of ectomycorrhizal fungi. Mol. Ecol. 25:630-647.

Looney BP, Ryberg M, Hampe F, Sánchez-García M, Matheny PB. 2015. Data from: Into and out of the tropics: global diversification patterns in a hyper-diverse clade of ectomycorrhizal fungi. In: Dryad.

Magallon S, Sanderson MJ. 2001. Absolute diversification rates in angiosperm clades. Evolution 55:1762-1780.

Martijn J, Vosseberg J, Guy L, Offre P, Ettema TJ. 2018. Deep mitochondrial origin outside the sampled alphaproteobacteria. Nature.

McCormack JE, Harvey MG, Faircloth BC, Crawford NG, Glenn TC, Brumfield RT. 2013a. Data from: A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. In: Dryad Data Repository.

McCormack JE, Harvey MG, Faircloth BC, Crawford NG, Glenn TC, Brumfield RT. 2013b. A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. PLoS One 8:e54848.

McKinney W. 2010. Data Structures for Statistical Computing in Python.

Meiklejohn KA, Faircloth BC, Glenn TC, Kimball RT, Braun EL. 2016a. Analysis of a Rapid Evolutionary Radiation Using Ultraconserved Elements: Evidence for a Bias in Some Multispecies Coalescent Methods. Syst. Biol. 65:612-627.

Meiklejohn KA, Faircloth BC, Glenn TC, Kimball RT, Braun EL. 2016b. Data from: Analysis of a rapid evolutionary radiation using ultraconserved elements (UCEs): Evidence for a bias in some multi-species coalescent methods. In: Dryad.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. 37:1530-1534.

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, et al. 2014a. Data from: Phylogenomics resolves the timing and pattern of insect evolution. In: Dryad Digital Repository.

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, et al. 2014b. Phylogenomics resolves the timing and pattern of insect evolution. Science 346:763-767.

Mooers AO, Holmes EC. 2000. The evolution of base composition and phylogenetic inference. Trends Ecol. Evol. 15:365-369.

Moyle RG, Oliveros CH, Andersen MJ, Hosner PA, Benz BW, Manthey JD, Travers SL, Brown RM, Faircloth BC. 2016a. Data from: Tectonic collision and uplift of Wallacea triggered the global songbird radiation. In: Dryad Data Repository.

Moyle RG, Oliveros CH, Andersen MJ, Hosner PA, Benz BW, Manthey JD, Travers SL, Brown RM, Faircloth BC. 2016b. Tectonic collision and uplift of Wallacea triggered the global songbird radiation. Nat Commun 7:12709.

Murray EA, Carmichael AE, Heraty JM. 2013a. Ancient host shifts followed by host conservatism in a group of ant parasitoids. Proc Biol Sci 280:20130495.

Murray EA, Carmichael AE, Heraty JM. 2013b. Data from: Ancient host shifts followed by host conservatism in a group of ant parasitoids. In: Dryad Data Repository.

Naser-Khdour S, Minh BQ, Zhang W, Stone EA, Lanfear R. 2019. The Prevalence and Impact of Model Violations in Phylogenetic Analysis. Genome Biol Evol.

Near TJ, Dornburg A, Eytan RI, Keck BP, Smith WL, Kuhn KL, Moore JA, Price SA, Burbrink FT, Friedman M, et al. 2013a. Data from: Phylogeny and tempo of diversification in the superradiation of spiny-rayed fishes. In: Dryad.

Near TJ, Dornburg A, Eytan RI, Keck BP, Smith WL, Kuhn KL, Moore JA, Price SA, Burbrink FT, Friedman M, et al. 2013b. Phylogeny and tempo of diversification in the superradiation of spiny-rayed fishes. Proc Natl Acad Sci U S A 110:12738-12743.

Nee S, May RM, Harvey PH. 1994. The reconstructed evolutionary process. Philos Trans R Soc Lond B Biol Sci 344:305-311.

Nguyen AD, Gotelli NJ, Cahan SH. 2016a. Data from: The evolution of heat shock protein sequences, cis-regulatory elements, and expression profiles in the eusocial Hymenoptera. In: Dryad.

Nguyen AD, Gotelli NJ, Cahan SH. 2016b. The evolution of heat shock protein sequences, cis-regulatory elements, and expression profiles in the eusocial Hymenoptera. BMC Evol. Biol. 16:15.

Nguyen LT, 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.

Oaks JR. 2011a. Data from: A time-calibrated species tree of Crocodylia reveals a recent radiation of the true crocodiles. In: Dryad Data Repository.

Oaks JR. 2011b. A time-calibrated species tree of Crocodylia reveals a recent radiation of the true crocodiles. Evolution 65:3285-3297.

Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21:1455-1458.

Phillips MJ, Penny D. 2003. The root of the mammalian tree inferred from whole mitochondrial genomes. Mol Phylogenet Evol 28:171-185.

Prebus M. 2017a. Data from: Insights into the evolution, biogeography and natural history of the acorn ants, genus Temnothorax Mayr (Hymenoptera: Formicidae). In: Dryad.

Prebus M. 2017b. Insights into the evolution, biogeography and natural history of the acorn ants, genus Temnothorax Mayr (hymenoptera: Formicidae). BMC Evol. Biol. 17:250.

Preparata G, Saccone C. 1987. A simple quantitative model of the molecular clock. J. Mol. Evol. 26:7-15.

Puttick MN, Morris JL, Williams TA, Cox CJ, Edwards D, Kenrick P, Pressel S, Wellman CH, Schneider H, Pisani D, et al. 2018. The Interrelationships of Land Plants and the Nature of the Ancestral Embryophyte. Curr. Biol. 28:733-745 e732.

Pyron RA, Wiens JJ. 2011. A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. Mol Phylogenet Evol 61:543-583.

Pyron RA, Wiens JJ, Alexander Pyron R. 2011. Data from: A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. In: Dryad.

Ran JH, Shen TT, Wang MM, Wang XQ. 2018a. Data from: Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between Gnetales and angiosperms. In: Dryad Digital Repository.

Ran JH, Shen TT, Wang MM, Wang XQ. 2018b. Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between Gnetales and angiosperms. Proc Biol Sci 285:20181012.

Rannala B, Yang Z. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J. Mol. Evol. 43:304-311.

Reddy S, Kimball RT, Pandey A, Hosner PA, Braun MJ, Hackett SJ, Han K, Harshman J, Huddleston CJ, Kingston S, et al. 2017a. Data from: Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. In: Dryad Digital Repository.

Reddy S, Kimball RT, Pandey A, Hosner PA, Braun MJ, Hackett SJ, Han KL, Harshman J, Huddleston CJ, Kingston S, et al. 2017b. Why Do Phylogenomic Data Sets Yield Conflicting Trees? Data Type Influences the Avian Tree of Life more than Taxon Sampling. Syst. Biol. 66:857-879.

Richart CH, Hayashi CY, Hedin M. 2016a. Data from: Phylogenomic analyses resolve an ancient trichotomy at the base of Ischyropsalidoidea (Arachnida, Opiliones) despite high levels of gene tree conflict and unequal minority resolution frequencies. In: Dryad.

Richart CH, Hayashi CY, Hedin M. 2016b. Phylogenomic analyses resolve an ancient trichotomy at the base of Ischyropsalidoidea (Arachnida, Opiliones) despite high levels of gene tree conflict and unequal minority resolution frequencies. Mol Phylogenet Evol 95:171-182.

Rightmyer MG, Griswold T, Brady SG. 2013a. Data from: Phylogeny and systematics of the bee genus Osmia (Hymenoptera: Megachilidae) with emphasis on North American Melanosmia: subgenera, synonymies, and nesting biology revisited. In: Dryad Data Repository.

Rightmyer MG, Griswold T, Brady SG. 2013b. Phylogeny and systematics of the bee genus Osmia (Hymenoptera: Megachilidae) with emphasis on North American Melanosmia: subgenera, synonymies and nesting biology revisited. Syst. Entomol. 38:561-576.

Roberts D, Yang Z. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol. Biol. Evol. 12:451-458.

Robinson DF, Foulds LR. 1981. Comparison of Phylogenetic Trees. Math. Biosci. 53:131-147.

Sauquet H, Ho SY, Gandolfo MA, Jordan GJ, Wilf P, Cantrill DJ, Bayly MJ, Bromham L, Brown GK, Carpenter RJ, 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.

Sauquet H, Ho SYW, Gandolfo MA, Jordan GJ, Wilf P, Cantrill DJ, Bayly MJ, Bromham L, Brown GK, Carpenter RJ, et al. 2011. Data from: Testing the impact of calibration on molecular divergence times using a fossil-rich group: the case of Nothofagus (Fagales). In: Dryad Data Repository.

Seago AE, Giorgi JA, Li J, Ślipiński A. 2011a. Data from: Phylogeny, classification and evolution of ladybird beetles (Coleoptera: Coccinellidae) based on simultaneous analysis of molecular and morphological data. In: Dryad Data Repository.

Seago AE, Giorgi JA, Li J, Ślipiński A. 2011b. Phylogeny, classification and evolution of ladybird beetles (Coleoptera: Coccinellidae) based on simultaneous analysis of molecular and morphological data. Mol. Phylogen. Evol. 60:137-151.

Sharanowski BJ, Dowling APG, Sharkey MJ. 2011a. Data from: Molecular phylogenetics of Braconidae (Hymenoptera: Ichneumonoidea) based on multiple nuclear genes and implications for classification. In: Dryad Data Repository.

Sharanowski BJ, Dowling APG, Sharkey MJ. 2011b. Molecular phylogenetics of Braconidae (Hymenoptera: Ichneumonoidea), based on multiple nuclear genes, and implications for classification. Syst. Entomol. 36:549-572.

Shen X-X. 2018. Data from: Tempo and mode of genome evolution in the budding yeast subphylum. In: Figshare.

Shen XX, Opulente DA, Kominek J, Zhou X, Steenwyk JL, Buh KV, Haase MAB, Wisecaver JH, Wang M, Doering DT, et al. 2018. Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. Cell 175:1533-1545 e1520.

Siler C, Brown RM, Oliveros CH, Santanen A. 2013. Data from: Multilocus phylogeny reveals unexpected diversification patterns in Asian Wolf Snakes (genus Lycodon). In: Dryad Data Repository.

Siler CD, Oliveros CH, Santanen A, Brown RM. 2013. Multilocus phylogeny reveals unexpected diversification patterns in Asian wolf snakes (genus Lycodon). Zool. Scr. 42:262-277.

Smith BT, Harvey MG, Faircloth BC, Glenn TC, Brumfield RT. 2014a. Data from: Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. In: Dryad Digital Repository.

Smith BT, Harvey MG, Faircloth BC, Glenn TC, Brumfield RT. 2014b. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. Syst. Biol. 63:83-95.

Song N, Lin A, Zhao X. 2018. Insight into higher-level phylogeny of Neuropterida: Evidence from secondary structures of mitochondrial rRNA genes and mitogenomic data. PLoS One 13:e0191826.

Stadler T. 2013. How can we improve accuracy of macroevolutionary rate estimates? Syst. Biol. 62:321-329.

Stadler T. 2011. Simulating Trees with a Fixed Number of Extant Species. Syst. Biol. 60:676-684.

Steel MA, Penny D. 1993. Distributions of Tree Comparison Metrics - Some New Results. Syst. Biol. 42:126-141.

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. 1996. Phylogenetic Inference. In. Molecular systematics: Sunderland, Mass.: Sinauer Associates. p. 407-514.

Tamura K, Kumar S. 2002. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. Mol. Biol. Evol. 19:1727-1736.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10:512-526.

Tavaré S. 1986. Some probabilistic and statistical probles in the analysis of DNA sequences. Lectures on Mathematics in the Life Sciences 17.

Nguyen MAT, Gesell T, von Haeseler A. 2012. ImOSM: intermittent evolution and robustness of phylogenetic methods. Mol. Biol. Evol. 29:663-673.

Tolley KA, Townsend TM, Vences M. 2013a. Data from: Large-scale phylogeny of chameleons suggests African origins and Eocene diversification. In: Dryad Data Repository.

Tolley KA, Townsend TM, Vences M. 2013b. Large-scale phylogeny of chameleons suggests African origins and Eocene diversification. Proc Biol Sci 280:20130184.

Unmack PJ, Allen GR, Johnson JB. 2013a. Data from: Phylogeny and biogeography of rainbowfishes (Melanotaeniidae) from Australia and New Guinea. In: Dryad Data Repository.

Unmack PJ, Allen GR, Johnson JB. 2013b. Phylogeny and biogeography of rainbowfishes (Melanotaeniidae) from Australia and New Guinea. Mol Phylogenet Evol 67:15-27.

Varga T, Krizsán K, Földi C, Dima B, Sánchez-García M, Sánchez-Ramírez S, Szöllősi GJ, Szarkándi JG, Papp V, Albert L, et al. 2019a. Data from: Megaphylogeny resolves global patterns of mushroom evolution. In: Dryad.

Varga T, Krizsán K, Földi C, Dima B, Sánchez-García M, Sánchez-Ramírez S, Szöllősi GJ, Szarkándi JG, Papp V, Albert L, et al. 2019b. Megaphylogeny resolves global patterns of mushroom evolution. Nat Ecol Evol 3:668-678.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17:261-272.

von Haeseler A, Janke A, Pääbo S. 1993. Molecular phylogenetics. Verhandlungen der Deutschen Zoologischen Gesellschaft= Proceedings of the German Zoological Society 86:119-129.

Wainwright PC, Smith WL, Price SA, Tang KL, Sparks JS, Ferry LA, Kuhn KL, Near TJ. 2012. Data from: The evolution of pharyngognathy: a phylogenetic and functional appraisal of the pharyngeal jaw key innovation in labroid fishes and beyond. In: Dryad Data Repository.

Wainwright PC, Smith WL, Price SA, Tang KL, Sparks JS, Ferry LA, Kuhn KL, Near TJ, Eytan RI. 2012. The evolution of pharyngognathy: a phylogenetic and functional appraisal of the pharyngeal jaw key innovation in labroid fishes and beyond. Syst. Biol. 61:1001-1027.

Walt Svd, Colbert SC, Varoquaux G. 2011. The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science & Engineering 13:22-30.

Weiss G, von Haeseler A. 2003. Testing Substitution Models Within a Phylogenetic Tree. Mol. Biol. Evol. 20:572-578.

Welch BL. 1947. The generalization of ‘STUDENT'S’problem when several different population varlances are involved. Biometrika 34:28-35.

Whelan NV, Kocot KM, Moroz TP, Mukherjee K, Williams P, Paulay G, Moroz LL, Halanych KM. 2017a. Author Correction: Ctenophore relationships and their placement as the sister group to all other animals. Nat Ecol Evol 1:1783.

Whelan NV, Kocot KM, Moroz TP, Mukherjee K, Williams P, Paulay G, Moroz LL, Halanych KM. 2017b. Data from: Ctenophora Phylogeny Datasets and Core Orthologs. In: Figshare.

Wood HM, Matzke NJ, Gillespie RG, Griswold CE. 2012. Data from: Treating fossils as terminal taxa in divergence time estimation reveals ancient vicariance patterns in the palpimanoid spiders. In: Dryad Data Repository.

Wood HM, Matzke NJ, Gillespie RG, Griswold CE. 2013. Treating fossils as terminal taxa in divergence time estimation reveals ancient vicariance patterns in the palpimanoid spiders. Syst. Biol. 62:264-284.

Worobey M, Han G, Rambaut A. 2014a. Data from: A synchronized global sweep of the internal genes of modern avian influenza virus. In: Dryad Data Repository.

Worobey M, Han GZ, Rambaut A. 2014b. A synchronized global sweep of the internal genes of modern avian influenza virus. Nature 508:254-257.

Wu S, Edwards S, Liu L. 2019. Data from: Genome-scale DNA sequence data and the evolutionary history of placental mammals. In: Figshare.

Wu S, Edwards S, Liu L. 2018. Genome-scale DNA sequence data and the evolutionary history of placental mammals. Data Brief 18:1972-1975.

Yang Z. 2006. Computational Molecular Evolution. Oxford, UNITED KINGDOM: Oxford University Press USA - OSO.

Yang Z. 1994. Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39:105-111.

Yang Z. 1995. A space-time process model for the evolution of DNA sequences. Genetics 139:993.

Zou L, Susko E, Field C, Roger AJ. 2012. Fitting nonstationary general-time-reversible models to obtain edge-lengths and frequencies for the Barry–Hartigan model. Syst. Biol. 61:927-940.

Appendix

TABLE A.1. Number of Taxa, Number of Sites, Type, Clade, and Study Reference for each dataset that has been used in this study

	Dataset	Study Reference	Dataset Reference	Type	Clade	Taxa	Sites	Partitions
1	Anderson_2013	(Anderson, et al. 2014)	(Anderson, et al. 2013)	DNA	Loliginidae	145	3037	4
2	Ballesteros_2019	(Ballesteros and Sharma 2019a)	(Ballesteros and Sharma 2019b)	AA	Chelicerata	53	1484206	3534
3	Becker_2016	(Becker, et al. 2016)	(Becker, et al. 2017)	AA	Halobacteriacea	170	217	1
4	Bergsten _2013	(Bergsten, et al. 2013a)	(Bergsten, et al. 2013b)	DNA	Dytiscidae	38	2111	8
5	Borowiec_2015	(Borowiec, et al. 2015)	(Borowiec, et al. 2016)	AA	Metazoa	36	384981	1080
6	Branstetter_2017	(Branstetter, et al. 2017b)	(Branstetter, et al. 2017a)	DNA	Aculeata	187	183747	807
7	Broughton_2013	(Broughton, et al. 2013b)	(Broughton, et al. 2013a)	DNA	Osteichthyes	61	19997	61
8	Brown_2012	(Brown, et al. 2012b)	(Brown, et al. 2012a)	DNA	Ptychozoon	41	1665	7
9	Cannon_2016a	(Cannon, et al. 2016a)	(Cannon, et al. 2016b)	AA	Metazoa	78	44896	212
10	Cannon_2016b	(Cannon, et al. 2016a)	(Cannon, et al. 2016b)	DNA	Metazoa	78	89792	424
11	Chen_2015	(Chen, et al. 2015b)	(Chen, et al. 2015a)	AA	Gnathostomata	58	1806035	4682
12	Cognato_2001	(Cognato and Vogler 2001b)	(Cognato and Vogler 2001a)	DNA	Coleoptera: Scolytinae	44	1897	7
13	Crawford_2012	(Crawford, et al. 2012b)	(Crawford, et al. 2012a)	DNA	Sauria	10	465241	1145
14	Day_2013	Day, et al. (2013a)	(Day, et al. 2013b)	DNA	Synodontis	152	3586	11
15	Devitt_2013	(Devitt, Devitt, et al. 2013)	(Devitt, Cameron Devitt, et al. 2013)	DNA	Ensatina eschscholtzii klauberi	69	823	4
16	Dornburg_2012	(Dornburg, et al. 2012b)	(Dornburg, et al. 2012a)	DNA	Teleostei: Beryciformes: Holocentridae	44	5919	21
17	Faircloth_2013	(Faircloth, et al. 2013b)	(Faircloth, et al. 2013a)	DNA	Actinopterygii	27	149366	491
18	Fong_2012	(Fong, et al. 2012b)	(Fong, et al. 2012a)	DNA	Vertebrata	110	25919	168
19	Horn_2014	(Horn, et al. 2014b)	(Horn, et al. 2014a)	DNA	Euphorbia	197	11587	28
20	Irisarri_2017	(Irisarri, et al. 2017b)	(Irisarri, et al. 2017a)	AA	Gnathostomata	100	1964439	4593
21	Jarvis_2015	(Jarvis, et al. 2015)	(Jarvis, et al. 2014)	AA	Aves	52	4519041	8295

22	Kawahara_2013	(Kawahara and Rubinoff 2013a)	(Kawahara and Rubinoff 2013b)	DNA	Hyposmocoma	70	2238	9
23	Lartillot_2012	(Lartillot and Delsuc 2012b)	(Lartillot and Delsuc 2012a)	DNA	Eutheria	78	15117	51
24	Leache_2015	(Leache, et al. 2015)	(Leaché, et al. 2015)	DNA	Phrynosomatinae	11	358363	583
25	Looney_2016	(Looney, et al. 2016)	(Looney, et al. 2015)	DNA	Russula	1171	3927	4
26	McCormack_2013	(McCormack, et al. 2013b)	(McCormack, et al. 2013a)	DNA	Neoaves	33	539526	1541
27	Meiklejohn_2016	(Meiklejohn, et al. 2016a)	(Meiklejohn, et al. 2016b)	DNA	Phasianidae	18	614159	1501
28	Misof_2014	(Misof, et al. 2014b)	(Misof, et al. 2014a)	AA	Insecta	144	595033	2868
29	Moyle_2016	(Moyle, et al. 2016b)	(Moyle, et al. 2016a)	DNA	Oscines	106	375172	515
30	Murray_2013	(Murray, et al. 2013a)	(Murray, et al. 2013b)	DNA	Eucharitidae	237	3111	9
31	Near_2013	(Near, et al. 2013b)	(Near, et al. 2013a)	DNA	Acanthomorpha	608	8577	30
32	Nguyen_2016a	(Nguyen, et al. 2016b)	(Nguyen, et al. 2016a)	AA	Hymenoptera	17	688	1
33	Nguyen_2016b	(Nguyen, et al. 2016b)	(Nguyen, et al. 2016a)	AA	Hymenoptera	31	680	1
34	Nguyen_2016c	(Nguyen, et al. 2016b)	(Nguyen, et al. 2016a)	AA	Hymenoptera	25	811	1
35	Nguyen_2016d	(Nguyen, et al. 2016b)	(Nguyen, et al. 2016a)	AA	Hymenoptera	17	704	1
36	Nguyen_2016e	(Nguyen, et al. 2016b)	(Nguyen, et al. 2016a)	AA	Hymenoptera	17	385	1
37	Nguyen_2016f	(Nguyen, et al. 2016b)	(Nguyen, et al. 2016a)	AA	Hymenoptera	17	583	1
38	Oaks_2011	(Oaks 2011b)	(Oaks 2011a)	DNA	Crocodylia	79	7282	50
39	Prebus_2017	(Prebus 2017b)	(Prebus 2017a)	DNA	Temnothorax	50	1561581	2098
40	Pyron_2011	(Pyron and Wiens 2011)	(Pyron, et al. 2011)	DNA	Amphibia	2872	12712	34
41	Ran_2018a	(Ran, et al. 2018b)	(Ran, et al. 2018a)	AA	Spermatophyta	38	432014	1308
42	Ran_2018b	(Ran, et al. 2018b)	(Ran, et al. 2018a)	DNA	Spermatophyta	38	1296042	3924
43	Reddy_2017	(Reddy, et al. 2017b)	(Reddy, et al. 2017a)	DNA	Aves	235	137324	88
44	Richart_2015	(Richart, et al. 2016b)	(Richart, et al. 2016a)	DNA	Ischyropsalidoidea	6	536124	2016
45	Rightmyer_2013	(Rightmyer, et al. 2013b)	(Rightmyer, et al. 2013a)	DNA	Hymenoptera: Megachilidae	94	3692	25
46	Sauquet_2011	(Sauquet, et al. 2012)	(Sauquet, et al. 2011)	DNA	Nothofagus	51	5444	10
47	Seago_2011	(Seago, et al. 2011b)	(Seago, et al. 2011a)	DNA	Coccinellidae	97	2253	7
48	Sharanowski_2011	(Sharanowski, et al. 2011b)	(Sharanowski, et al. 2011a)	DNA	Braconidae	139	3982	11
49	Shen_2018	(Shen, et al. 2018)	(Shen 2018)	AA	Saccharomycotina	343	1162805	2407

50	Siler_2013	(Siler, Oliveros, et al. 2013)	(Siler, Brown, et al. 2013)	DNA	Lycodon	61	2697	7
51	Smith_2014	(Smith, et al. 2014b)	(Smith, et al. 2014a)	DNA	Xenops minutus	8	825804	1366
52	Tolley_2013	(Tolley, et al. 2013b)	(Tolley, et al. 2013a)	DNA	Chamaeleonidae	203	5054	16
53	Unmack_2013	(Unmack, et al. 2013b)	(Unmack, et al. 2013a)	DNA	Melanotaeniidae	139	6827	25
54	Varga_2019	(Varga, et al. 2019b)	(Varga, et al. 2019a)	DNA	Basidiomycota	5285	5737	3
55	Wainwright_2012	(Wainwright, et al. 2012a)	(Wainwright, et al. 2012b)	DNA	Acanthomorpha	188	8439	30
56	Whelan_2017	(Whelan, et al. 2017a)	(Whelan, et al. 2017b)	AA	Metazoa	76	49388	127
57	Wood_2012	(Wood, et al. 2013)	(Wood, et al. 2012)	DNA	Archaeidae	37	5185	8
58	Worobey_2014a	(Worobey, et al. 2014b)	(Worobey, et al. 2014a)	DNA	Influenzavirus A	146	1716	3
59	Worobey_2014b	(Worobey, et al. 2014b)	(Worobey, et al. 2014a)	DNA	Influenzavirus A	327	759	3
60	Worobey_2014c	(Worobey, et al. 2014b)	(Worobey, et al. 2014a)	DNA	Influenzavirus A	92	1416	3
61	Worobey_2014d	(Worobey, et al. 2014b)	(Worobey, et al. 2014a)	DNA	Influenzavirus A	355	1497	3
62	Worobey_2014e	(Worobey, et al. 2014b)	(Worobey, et al. 2014a)	DNA	Influenzavirus A	340	699	3
63	Worobey_2014f	(Worobey, et al. 2014b)	(Worobey, et al. 2014a)	DNA	Influenzavirus A	332	2151	3
64	Worobey_2014g	(Worobey, et al. 2014b)	(Worobey, et al. 2014a)	DNA	Influenzavirus A	326	2274	3
65	Worobey_2014h	(Worobey, et al. 2014b)	(Worobey, et al. 2014a)	DNA	Influenzavirus A	351	2280	3
66	Wu_2018a	(Wu, et al. 2018)	(Wu, et al. 2019)	AA	mammalia	90	3050199	5162
67	Wu_2018b	(Wu, et al. 2018)	(Wu, et al. 2019)	DNA	mammalia	90	9150597	15486

TABLE A.2. The probability distributions and their probability density function that we used for the Kolmogorov-Smirnov test.

distribution	PDF	notes
Alpha	$\frac{1}{x^2\Phi(\alpha)\sqrt{2\pi}}e^{-\frac{1}{2}(\alpha-\frac{1}{x})^2}$	Φ is the normal CDF α is the shape parameter
Beta	$\frac{\Gamma(\alpha+\beta)x^{\alpha-1}(1-x)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}$	α and β are the shape parameters Γ is the gamma function
Bradford	$\frac{\alpha}{\log(1+\alpha)(1+\alpha x)}$	α is the shape parameter
Chi	$\frac{1}{2^{\alpha/2-1}\Gamma(\frac{\alpha}{2})}x^{\alpha-1}e^{-\frac{x^2}{2}}$	α is the degrees of freedom Γ is the gamma function
Chi-squared	$\frac{1}{2^{\alpha/2}\Gamma(\frac{\alpha}{2})}x^{\alpha/2-1}e^{-\frac{x}{2}}$	α is the degrees of freedom Γ is the gamma function
Double gamma	$\frac{\alpha}{2\Gamma(\alpha)}\lvert x\rvert^{\alpha-1}e^{-\lvert x\rvert}$	α is the shape parameter Γ is the gamma function
Double Weibull	$\frac{\alpha}{2}\lvert x\rvert^{\alpha-1}e^{-\lvert x\rvert^{\alpha}}$	α is the shape parameter
Exponential normal	$\frac{1}{2\alpha}e^{\frac{1}{2\alpha^2}-\frac{x}{\alpha}}erfc(-\frac{x-\frac{1}{\alpha}}{\sqrt{2}})$	α is the shape parameter
Exponential Weibull	$\alpha\beta(1-e^{-x^{\beta}})^{\alpha-1}e^{-x^{\beta}}x^{\beta-1}$	α and β are the shape parameters
Exponential power	$\alpha x^{\alpha-1}e^{1+x^{\alpha}-e^{x^{\alpha}}}$	α is the shape parameter
Gamma		
Generalized logistic	$\alpha\frac{e^{-x}}{(1+e^{-x})^{\alpha+1}}$	α is the shape parameter
Generalized Pareto	$(1+\alpha x)^{-1-\frac{1}{\alpha}}$	α is the shape parameter
Generalized normal	$\frac{\alpha}{2\Gamma(1/\alpha)}e^{-\lvert x\rvert^{\alpha}}$	α is the shape parameter Γ is the gamma function
Generalized exponential	$(\alpha+\beta(1$ $-e^{\gamma x}))e^{-\alpha x-\beta x+\frac{\beta}{\gamma}(1-e^{-\gamma x})}$	α, β, γ are the shape parameters
Generalized gamma	$\frac{\lvert\beta\rvert x^{\alpha\beta-1}}{\Gamma(\alpha)}e^{-x^{\beta}}$	α and β are the shape parameters
Half-logistic	$\frac{2e^{-x}}{(1+e^{-x})^2}$	
Half-normal	$\sqrt{2/\pi}e^{\frac{-x^2}{2}}$	
Upped half of the generalized normal	$\frac{\alpha}{\Gamma(1/\alpha)}e^{-\lvert x\rvert^{\alpha}}$	α is the shape parameter
Inverse-gamma	$\frac{x^{-\alpha-1}}{\Gamma(\alpha)}e^{-\frac{1}{x}}$	α is the shape parameter

Inverse-normal	$\frac{1}{\sqrt{2\pi x^3}} e^{-\frac{(x-\alpha)^2}{2x\alpha^2}}$	α is the shape parameter
Inverse-Weibull	$\alpha x^{-\alpha-1} e^{-x^{-\alpha}}$	α is the shape parameter
Laplace	$\frac{1}{2} e^{-\lvert x\rvert}$	
Log-gamma	$\frac{e^{\alpha x - e^{x}}}{\Gamma(\alpha)}$	α is the shape parameter Γ is the gamma function
Logistic	$\frac{e^{-x}}{(1+e^{-x})^2}$	
Log-Laplace	$\frac{\alpha}{2} x^{\alpha-1} \quad 0 < x < 1$ $\frac{\alpha}{2} x^{-\alpha-1} \quad x \geq 1$	α is the shape parameter
Log-normal	$\frac{1}{\alpha x\sqrt{2\pi}} e^{-\frac{(\log x)^2}{2\alpha^2}}$	α is the shape parameter
Maxwell	$\sqrt{2/\pi} x^2 e^{\frac{-x^2}{2}}$	
Normal	$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$	
Pareto	$\frac{\alpha}{x^{\alpha+1}}$	α is the shape parameter
Power-law	$\alpha x^{\alpha-1}$	α is the shape parameter
Power Log-normal	$\frac{\alpha}{x\beta} \phi(\frac{\log x}{\beta})(\Phi\left(-\frac{\log x}{\beta}\right))^{\alpha-1}$	α and β are the shape parameters ϕ is the normal PDF Φ is the normal CDF
Power normal	$\alpha\phi(\mathrm{x})(\Phi(-\mathrm{x}))^{\alpha-1}$	α is the shape parameter ϕ is the normal PDF Φ is the normal CDF
Uniform	1	
Weibull maximum	$\alpha(-x)^{\alpha-1} e^{-(-x)^{\alpha}}$	α is the shape parameter
Weibull minimum	$\alpha x^{\alpha-1} e^{-x^{\alpha}}$	α is the shape parameter

Gamma function: $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx = (\alpha - 1)!$

TABLE A.3. False-negative rates for each of the MaxSym tests. The false-negative rates present the proportion of datasets that passed each MaxSym test when the parameter value is less than one.

			m=20	m=40	m=60	m=80	m=100	Mean ± SE
MaxSymTest	n=100	ρ	94.0	92.0	94.0	93.0	85.0	91.6
		ν	96.0	96.4	96.0	96.1	96.1	96.1
		ω	95.9	96.4	96.2	96.2	96.1	96.2
		ν + ω	95.9	96.4	95.9	96	95.9	96.0
		Δ	95.0	91.9	93.6	91.1	92.5	92.8
	n=1,000	ρ	71.0	49.0	49.0	37.0	43.0	49.8
		ν	85.5	83.9	85.7	83.4	82.5	84.2
		ω	86.4	84.9	85.9	83.2	82.1	84.5
		ν + ω	85.6	84.1	85.1	82.5	81.2	83.7
		Δ	62.9	62.2	64.3	64.4	70.2	64.8
	n=10,000	ρ	19.0	12.0	6.0	2.0	2.0	83.2
		ν	50.3	46	41.9	43.4	43.6	45.0
		ω	49.4	45.2	40.8	42.5	42.1	44.0
		ν + ω	47.9	43.2	38.3	40.1	40.4	42.0
		Δ	31.2	37.6	44.8	48.5	51.0	42.6
	Mean ± SE	ρ	61.3	51.0	49. 7	44.0	43.3	49.9±9.3
		ν	77.3	75.4	74.5	74.3	74.1	75.1±5.8
		ω	77.2	75.5	74.3	74.0	73.4	74.9±6.0
		ν + ω	76.5	74.6	73.1	72.9	72.5	73.9±6.2
		Δ	63.0	63.9	67.6	68.0	71.2	66.7±5.6
MaxSymTest$_{mar}$	n=100	ρ	96.0	93.0	89.0	89.0	89.0	91.2
		ν	94.4	95.4	95.2	94.5	95.3	95.0
		ω	93.7	95.4	95.6	94.5	95.4	94.9
		ν + ω	94.0	95.7	95.2	94.3	95.1	94.9
		Δ	92.4	88.7	91.3	90.4	91.6	90.9
	n=1,000	ρ	62.0	51.0	47.0	40.0	46.0	49.2
		ν	86.8	85.9	87.5	87.1	87.4	86.9
		ω	87.9	86.6	88.4	87.4	87.8	87.6
		ν + ω	87.0	85.6	87.5	86.7	87.1	86.8
		Δ	60.4	62.2	64.2	64.2	69.5	64.1
	n=10,000	ρ	20.0	19.0	8.0	4.0	3.0	10.8
		ν	58.2	60.5	62.0	63.6	63.4	61.5
		ω	60.0	62.7	63.7	65.5	65.6	63.5
		ν + ω	57.0	59.5	60.6	62.4	62.5	60.4
		Δ	31.9	37.8	44.9	48.2	51.2	42.8

	Mean ± SE	ρ	59.3	54.3	48.0	44.3	46.0	50.4±8.9
		ν	79.8	80.6	81.6	81.7	82.0	81.1±3.8
		ω	80.5	81.6	82.6	82.5	82.9	82.0±3.6
		ν + ω	79.3	80.3	81.1	81.1	81.6	80.7±3.9
		Δ	61.6	62.9	66.8	67.6	70.8	65.9±5.4
MaxSymTest$_{int}$	n=100	ρ	95.0	94.0	97.0	99.0	93.0	95.6
		ν	95.9	95.6	96.0	96.3	95.5	95.9
		ω	95.9	96.0	95.8	96.1	95.4	95.8
		ν + ω	95.8	95.8	95.9	96.1	95.2	95.8
		Δ	97.0	96.4	96.6	94.2	95.2	95.9
	n=1,000	ρ	89.0	87.0	92.0	88.0	87.0	88.6
		ν	90.8	91.0	90.2	88.5	87.7	89.6
		ω	91.0	90.9	90.0	88.0	87.2	89.4
		ν + ω	90.7	91.1	89.7	87.8	87.1	89.3
		Δ	91.8	89.9	91.8	91.8	92.4	91.5
	n=10,000	ρ	58.0	55.0	52.0	51.0	52.0	53.6
		ν	68.3	62.1	57.4	56.0	58.3	60.4
		ω	65.0	58.2	54.6	53.1	54.5	57.1
		ν + ω	65.6	58.4	53.8	52.5	54.8	57.0
		Δ	72.8	69.1	73.3	75.7	79.3	74.0
	Mean ± SE	ρ	80.7	78.7	80.3	79.3	77.3	79.3±4.9
		ν	85.0	82.9	81.2	80.3	80.5	82.0±4.2
		ω	84.0	81.7	80.1	79.1	79.0	80.8±4.6
		ν + ω	84.0	81.8	79.8	78.8	79.0	80.7±4.6
		Δ	87.2	85.1	87.2	87.2	89.0	87.2±2.6

TABLE A.4. False-positive rates for each of the MaxSym tests. The false-positive rates present the proportion of datasets that failed each MaxSym test when the parameter value equals to one.

			m=20	m=40	m=60	m=80	m=100	Mean ± SE
MaxSymTest	**n=100**	**ρ**	0.0	0.0	0.0	0.0	20.0	4.0
		ν	3.6	3.6	0.9	1.8	2.7	2.5
		ω	2.7	4.5	2.7	2.7	2.7	3.1
		ν + ω	0.0	10.0	0.0	0.0	10.0	4.0
	n=1,000	**ρ**	10.0	0.0	10.0	0.0	0.0	4.0
		ν	5.5	6.4	7.3	9.1	11.8	8.0
		ω	14.5	16.4	9.1	7.3	8.2	11.1
		ν + ω	10.0	0.0	20.0	0.0	10.0	8.0
	n=10,000	**ρ**	10.0	0.0	0.0	0.0	10.0	4.0
		ν	32.7	31.8	31.8	30.0	37.3	32.7
		ω	23.6	23.6	20.9	21.8	21.8	22.3
		ν + ω	0.0	0.0	10.0	0.0	0.0	2.0
	Mean ± SE	**ρ**	6. 7	0.0	3.3	0.0	10.0	4.0±1.6
		ν	13.9	13.9	13.3	13.6	17.3	14.4±3.6
		ω	13.6	14.8	10.9	10.6	10.9	12.2±2.2
		ν + ω	3.3	3.3	10.0	0.0	6.7	4.7±0.9
$\text{MaxSymTest}_{\text{mar}}$	**n=100**	**ρ**	0.0	0.0	10.0	0.0	10.0	4.0
		ν	8.2	8.2	0.9	3.6	1.8	4.5
		ω	0.9	8.2	5.5	3.6	2.7	4.2
		ν + ω	0.0	10.0	10.0	10.0	0.0	6.0
	n=1,000	**ρ**	0.0	0.0	20.0	0.0	0.0	4.0
		ν	2.7	2.7	3.6	6.4	6.4	4.4
		ω	13.6	10.0	12.7	9.1	10.9	11.3
		ν + ω	0.0	0.0	20.0	10.0	20.0	10.0
	n=10,000	**ρ**	0.0	10.0	0.0	0.0	0.0	2.0
		ν	10.0	4.5	4.5	3.6	2.7	5.1
		ω	28.2	26.4	21.8	21.8	25.5	24.7
		ν + ω	10.0	0.0	0.0	0.0	0.0	2.0
	Mean ± SE	**ρ**	0.0	3.3	10.0	0.0	3.3	3.3±1.6
		ν	7.0	5.1	3.0	4.5	3.6	4.7±0.7
		ω	14.2	14.9	13.3	11.5	13.0	13.4±2.4
		ν + ω	3.3	3.3	10.0	6.7	6.7	6.0±1.9
MaxSy mTest	**n=100**	**ρ**	0.0	10.0	20.0	0.0	20.0	10.0
		ν	2.7	1.8	4.5	5.5	3.6	3.6
		ω	2.7	5.5	2.7	3.6	2.7	3.4

	ν + ω	0.0	0.0	0.0	20.0	10.0	6.0
n=1,000	**ρ**	0.0	10.0	0.0	10.0	0.0	4.0
	ν	6.4	10.0	6.4	9.1	11.8	8.7
	ω	8.2	9.1	3.6	4.5	6.4	6.4
	ν + ω	10.0	0.0	0.0	0.0	10.0	4.0
n=10,000	**ρ**	0.0	0.0	0.0	10.0	10.0	4.0
	ν	37.3	39.1	34.5	37.3	43.6	38.4
	ω	4.5	0.9	6.4	8.2	6.4	5.3
	ν + ω	0.0	0.0	10.0	0.0	0.0	2.0
Mean ± SE	**ρ**	0.0	6. 7	6. 7	6. 7	10.0	6.0±1.9
	ν	15.5	17.0	15.1	17.3	19. 7	16.9±4.1
	ω	5.1	5.2	4.2	5.4	5.2	5.0±0.6
	ν + ω	3.3	0.0	3.3	6.7	6.7	4.0±1.6

TABLE A.5. False-negative rates for Chi-square test. The false-negative rates present the proportion of sequences that passed the chi-square test when the parameter value is less than one in the inheritance scheme simulations, and the difference between the two matrices is more than zero in the two-matrix scheme simulations.

			m=20	m=40	m=60	m=80	m=100	Mean ± SE
Chi-square Test	**n=100**	**ρ**	97.0	88.0	81.0	61.0	63.0	78.0
		ν	98.4	95.1	89.4	86.4	80.4	89.9
		ω	98.5	95.3	89.6	87.1	80.9	90.3
		ν + ω	98.5	95.2	89.3	86.3	80.1	89.9
		Δ	92.7	82.4	76.1	68.9	62.8	76.6
	n=1,000	**ρ**	62.0	32.0	14.0	12.0	6.0	25.2
		ν	84.1	72.7	62.9	59	52.4	66.2
		ω	85.6	75.1	65.7	61.7	55.3	68.7
		ν + ω	84.5	72.6	62.6	58.7	52.1	66.1
		Δ	58.7	45.6	40.0	34.9	33.7	42.6
	n=10,000	**ρ**	16.0	5.0	0.0	1.0	0.0	4.4
		ν	47.6	37.2	29.7	24.9	21.9	32.3
		ω	51.7	41.7	34.9	30.9	28.6	37.6
		ν + ω	47.0	36.1	29.1	24.9	22.3	31.9
		Δ	21.1	11.5	10.7	9.5	8.5	12.3
	Mean ± SE	**ρ**	58.3	41. 7	31. 7	24. 7	23.0	35.9±9.2
		ν	76.7	68.3	60. 7	56.8	51.6	62.8±6.8
		ω	78.6	70.7	63.4	59.9	54.9	65.5±6.2
		ν + ω	76.7	68.0	60.3	56.6	51.5	62.6±6.8
		Δ	57.5	46.5	42.3	37.8	35.0	43.8±7.4

TABLE A.6. False-positive rates for Chi2Cons test. The false-positive rates present the proportion of sequences that failed the chi-square test when the parameter value equals to one.

			m=20	m=40	m=60	m=80	m=100	Mean ± SE
Chi-square Test	n=100	ρ	0.0	0.0	0.0	0.0	10.0	2.0
		ν	0.9	5.5	7.3	6.4	10.0	6.0
		ω	2.7	7.3	9.1	13.6	15.5	9.6
		ν + ω	0.0	20.0	10.0	20.0	0.0	10.0
	n=1,000	ρ	0.0	0.0	10.0	20.0	30.0	12.0
		ν	2.7	0.9	2.7	8.2	13.6	5.6
		ω	17.3	24.5	30.9	35.5	42.7	30.2
		ν + ω	0.0	10.0	0.0	10.0	20.0	8.0
	n=10,000	ρ	0.0	0.0	0.0	0.0	20.0	4.0
		ν	0.9	2.7	6.4	8.2	9.1	5.5
		ω	41.8	47.3	58.2	68.2	76.4	58.4
		ν + ω	0.0	0.0	0.0	0.0	20.0	4.0
	Mean ± SE	ρ	0.0	0.0	3.3	6. 7	20.0	6.0 ± 2.5
		ν	1.5	3.0	5.5	7.6	10.9	5.7 ± 1.0
		ω	20.6	26.4	32.7	39.1	44.9	32.7 ± 5.9
		ν + ω	0.0	10.0	3.3	10.0	13.3	6.5 ± 2.0

TABLE A.7. False-negative rates for WH test. The false-negative rates present the proportion of datasets that passed the WH test when the parameter value is less than one.

			m=20	m=40		m=60	m=80	m=100	Mean ± SE
WH Test	**n=100**	**ρ**	100.0	100.0		100.0	100.0	NaN	100.0
		ν	79.4	84.0		66.7	88.9	100.0	83.8
		ω	77.4	83.3		66.7	88.9	87.5	80.8
		ν + ω	74.1	82.6		61.5	87.5	100.0	81.1
		Δ	99.9	99.9	99.7	99.7	99.8	99.8	
	n=1,000	**ρ**	8.7	1.9		2.0	0.0	0.0	2.5
		ν	16.9	13.0		10.7	10.3	10.7	12.3
		ω	8.2	2.3		0.7	0.7	0.4	2.5
		ν + ω	8.0	2.3		0.6	0.8	0.3	2.4
		Δ	62.1	60.6		60.3	61.9	65.8	62.1
	n=10,000	**ρ**	1.5	0.0		0.0	0.0	0.0	0.3
		ν	8.7	7.7		7.5	8.8	8.0	8.1
		ω	0.2	0.0	0.0	0.0	0.0	0.0	
		ν + ω	0.2	0.0		0.0	0.0	0.0	0.0
		Δ	37.4	39.1	40.8	40.1	39.5	39.4	
	Mean ± SE	**ρ**	36.7	34.0		34.0	33.3	0.0	29.5±12.4
		ν	35.0	34.9		28.3	36.0	39.6	34.8±9.4
		ω	27.4	28.3		20.7	29.4	33.4	27.6±10.1
		ν + ω	27.4	28.3		20.7	29.4	33.4	27.9±10.3
		Δ	66.5	66.5		66.9	67.2	68.4	67.1±6.7

TABLE A.8. False-positive rates for WH test. The false-positive rates present the proportion of datasets that failed the WH test when the parameter value equals to one.

		m=20	m=40		m=60	m=80	m=100	Mean ± SE
n=100	**ρ**	NaN	NaN	NaN	NaN	0.0	0.0	
	ν	0.0	0.0		0.0	0.0	50	10.0
	ω	0.0	0.0		0.0	0.0	0.0	0.0
	ν + ω	NaN	0.0		NaN	NaN	NaN	0.0
n=1,000	**ρ**	12.5	0.0	0.0	0.0	0.0	2.5	
	ν	79.5	87.8		90.6	89.7	91.8	87.9
	ω	2.3	2.2		2.3	7.0	1.1	3.0
	ν + ω	0.0	0.0		0.0	0.0	0.0	0.0
n=10,000	**ρ**	20.0	0.0	0.0	10.0	0.0	6.0	
	ν	93.6	90.9		90.9	91.8	90.9	91.6
	ω	13.7	21.2		21.6	14.2	19.2	18.0
	ν + ω	30.0	0.0		0.0	10.0	0.0	8.0
Mean ± SE	**ρ**	16.3	0.0	0.0	5.0	0.0	3.9 ± 2.1	
	ν	57.7	59.6		60.5	60.5	77.6	63.2 ± 10.5
	ω	5.3	7.8		8.0	7.1	6.8	6.8 ± 2.2
	ν + ω	15.0	0.0		0.0	5.0	0.0	3.6 ± 2.8

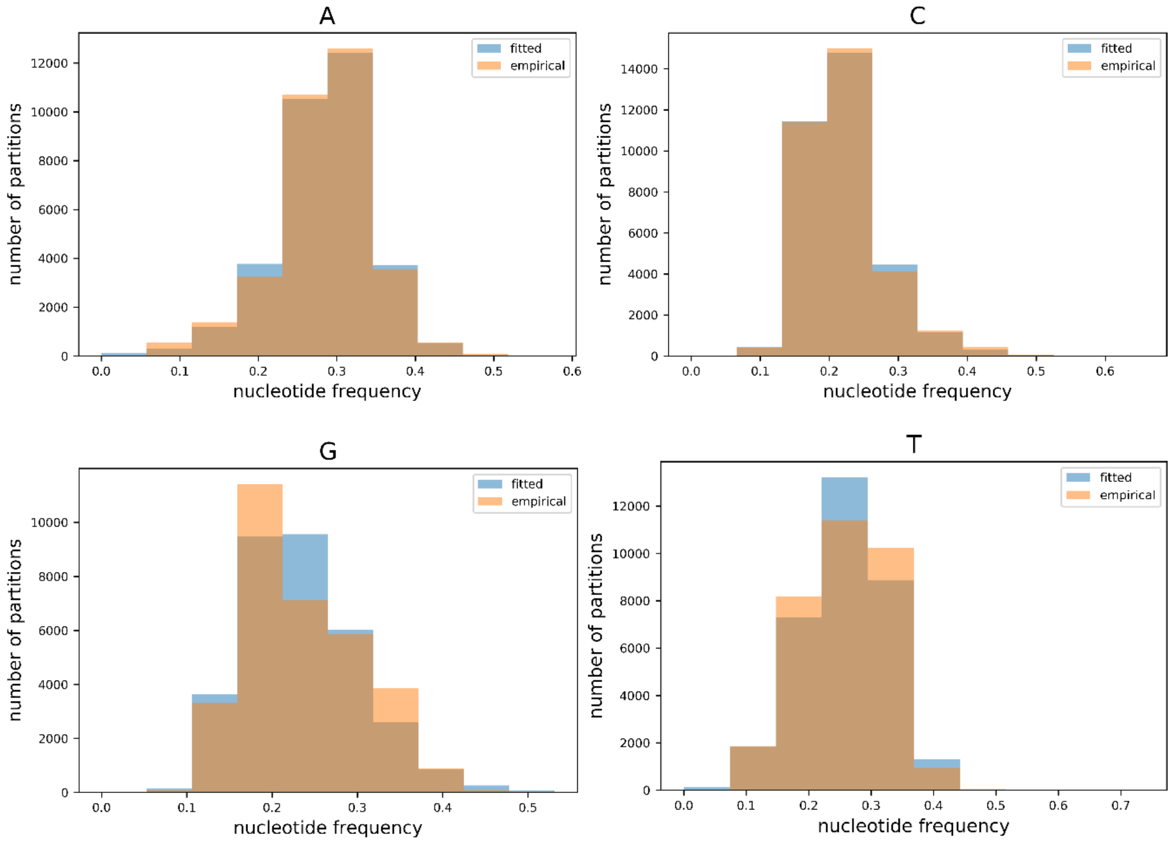


FIGURE A.1. distribution of the nucleotide frequencies. The empirical nucleotide frequencies for each single partition were estimated using IQ-TREE (orange) and the Fitted distribution (blue) were sampled from the best-fit distribution with the same number of partitions.

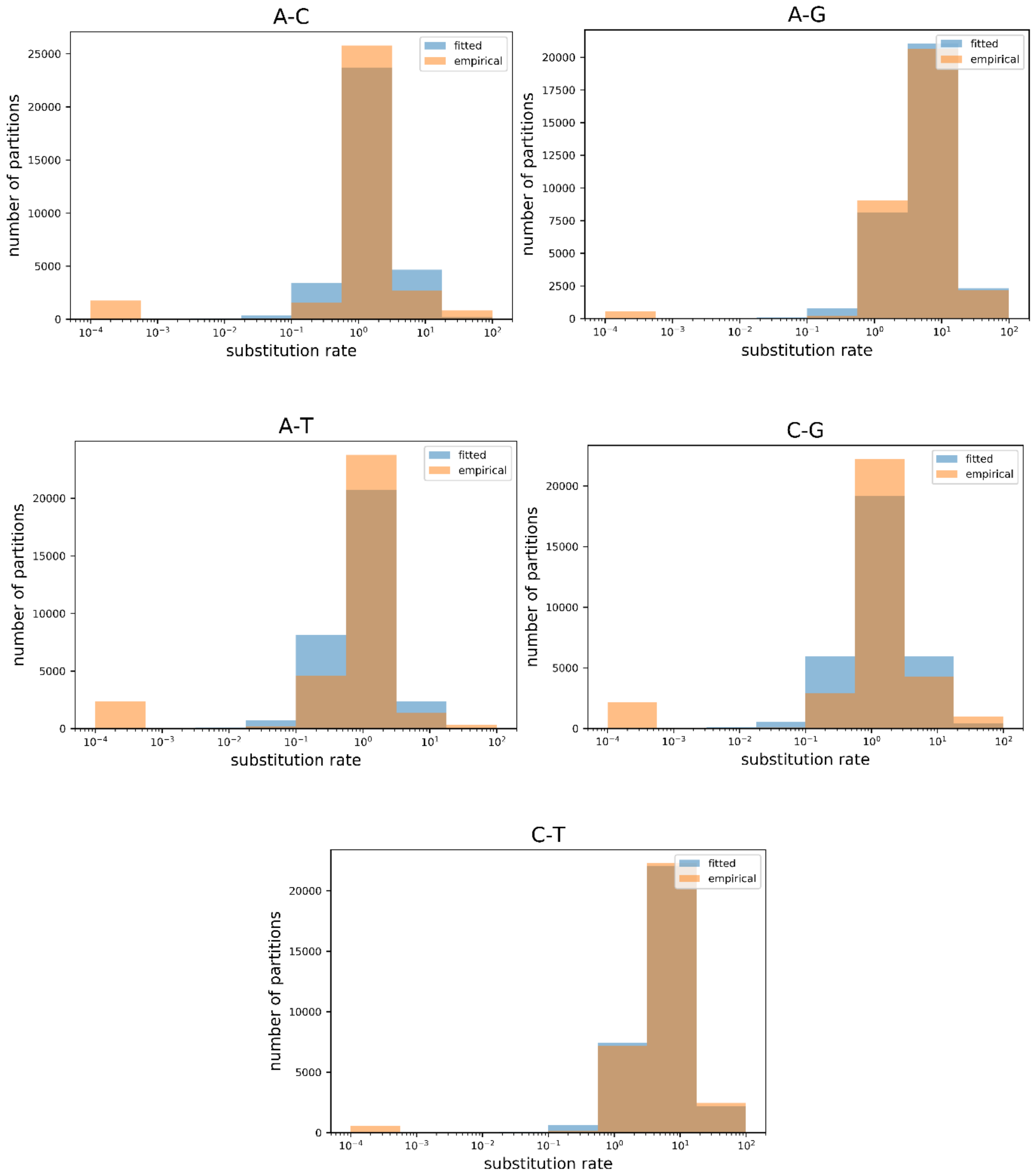


FIGURE A.2. distribution of the GTR parameters. The best-fit substitution rate matrix for each single partition was estimated using IQ-TREE (orange) and the Fitted distribution (blue) were sampled from the best-fit distribution with the same number of partitions.

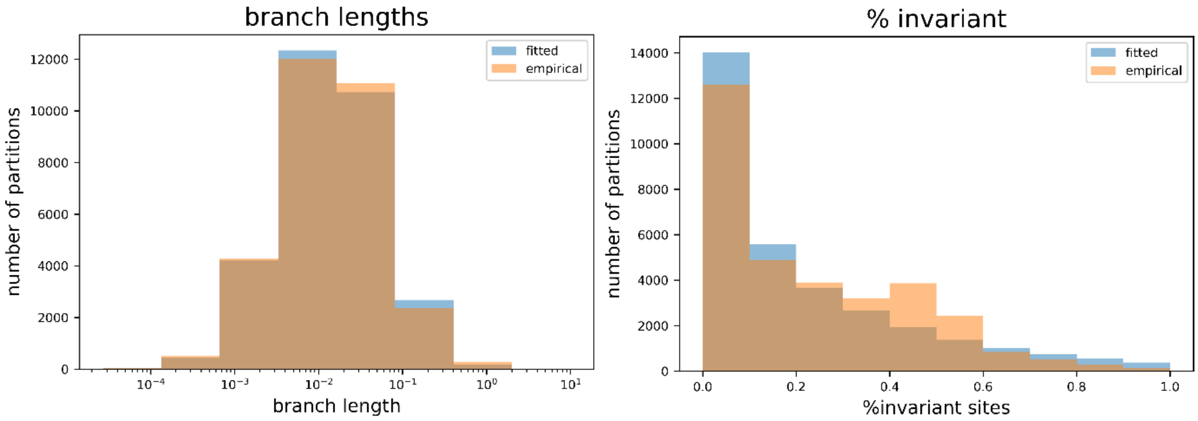


FIGURE A.3. The distribution of branch lengths and proportion of invariant sites

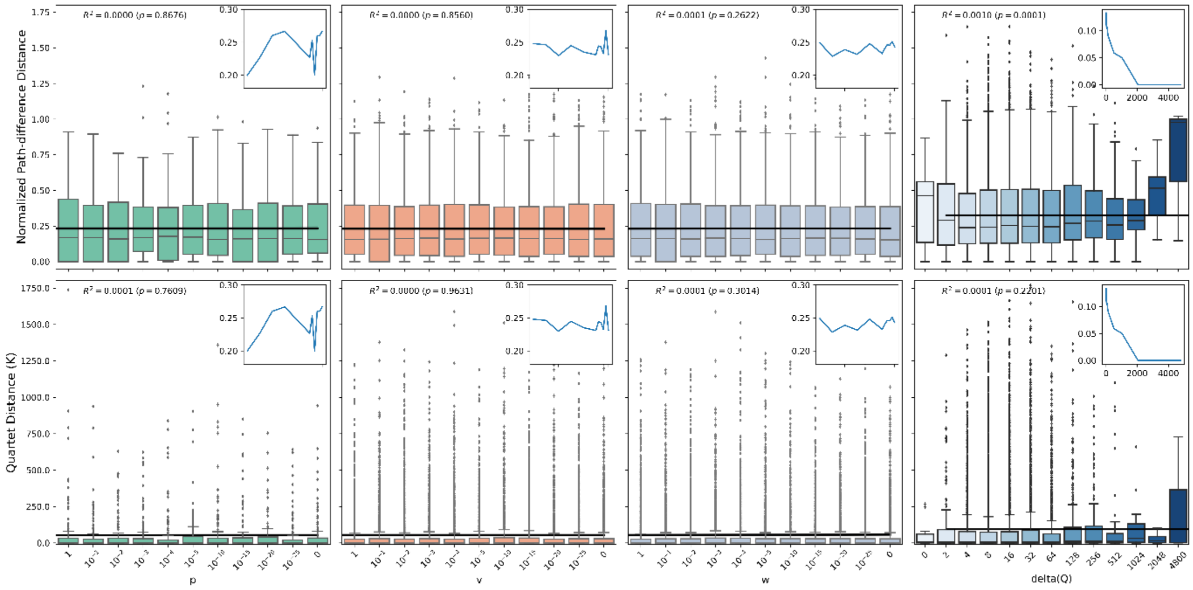


FIGURE A.4. Normalized Path-difference and Quartet metrics between the estimated tree topology and the original tree topology as a function of the inheritance weight (ν, ω ,ρ) and the distance between the two matrices. The small plots show the proportion of datasets in which the distance between the estimated topology and the original topology equals to zero as a function of the inheritance weight and the distance between the two matrices.

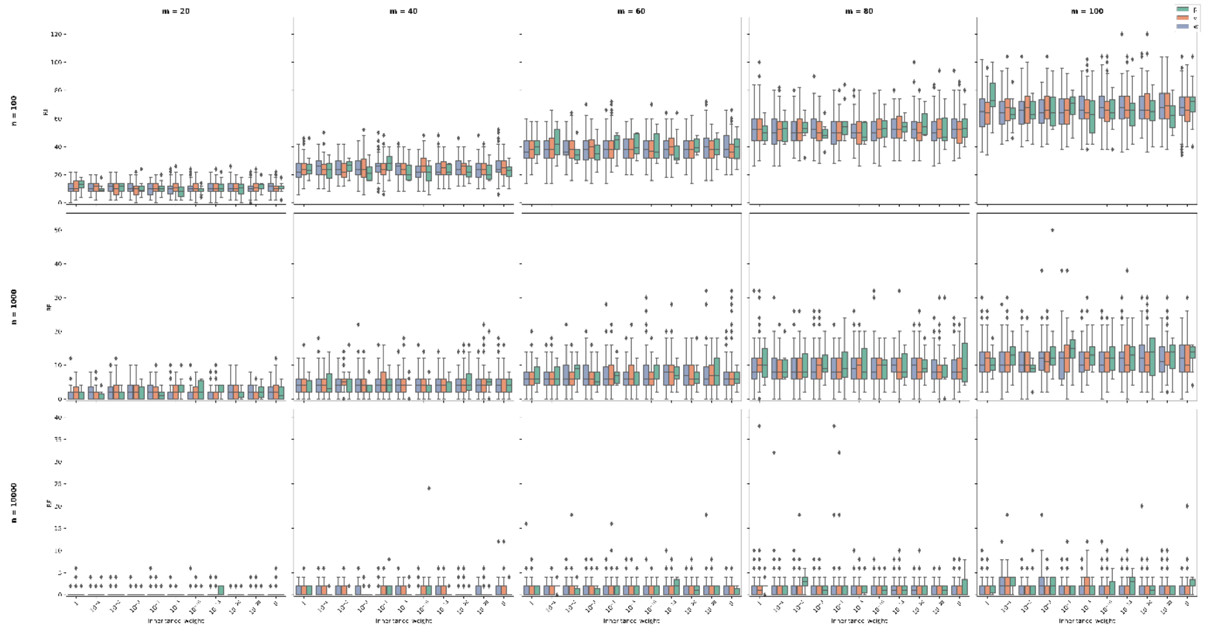


FIGURE A.5. The Robinson-Foulds metric as a function of the inheritance weight for each number of taxa (m) and number of site (n).

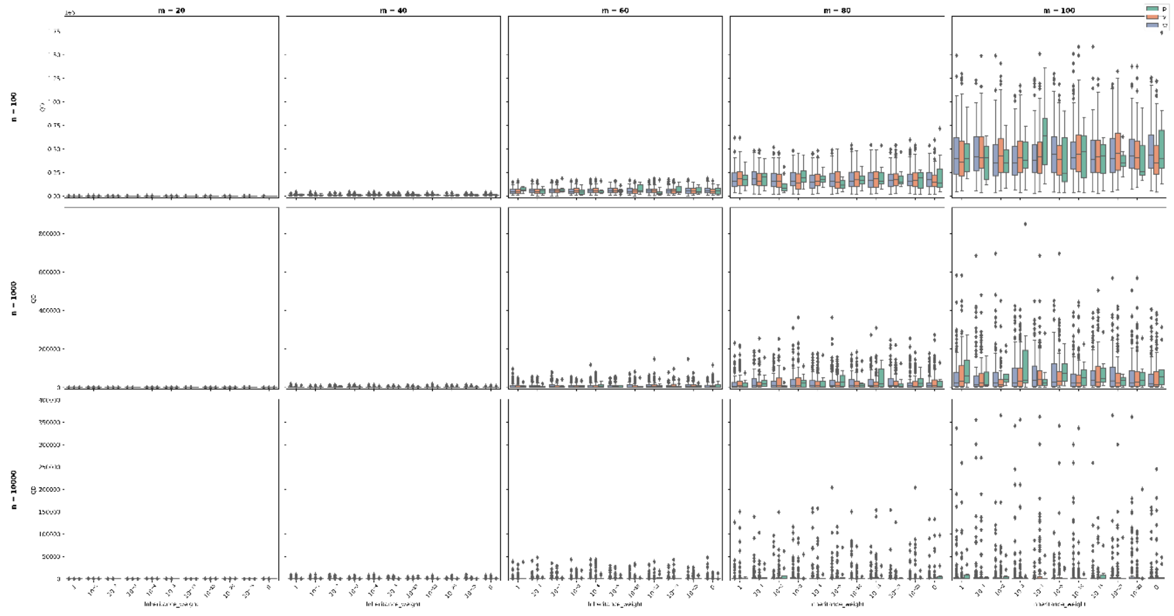


FIGURE A.6. The Quartet distance as a function of the inheritance weight for each number of taxa (m) and number of site (n).

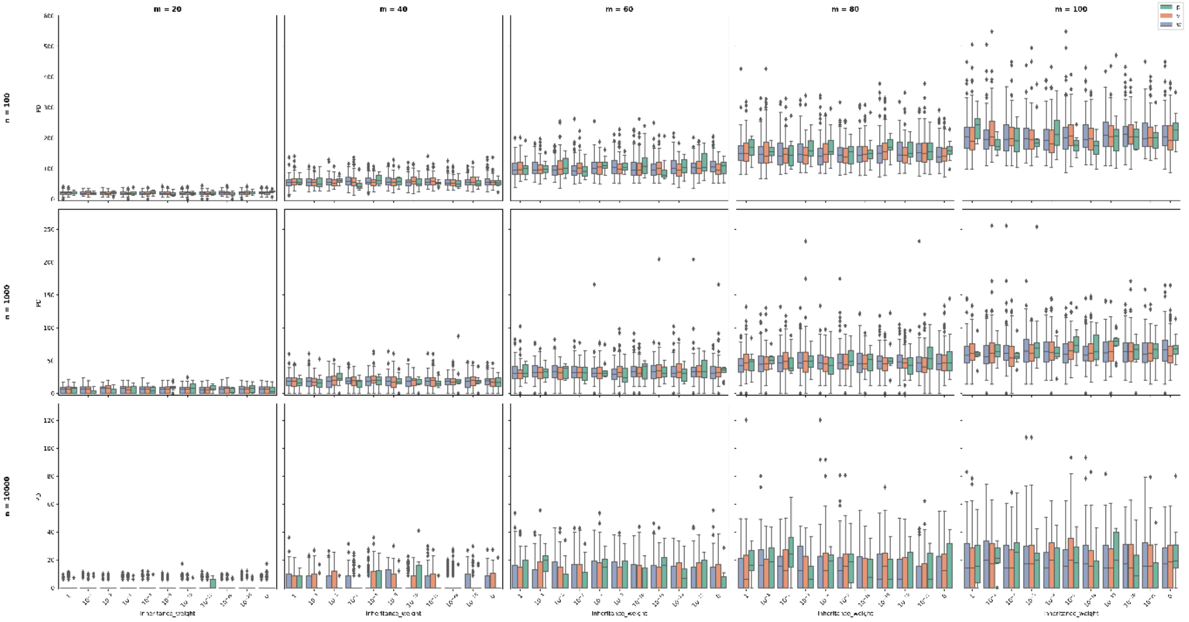


FIGURE A.7. The Path-Difference distance as a function of the inheritance weight for each number of taxa (m) and number of site (n).

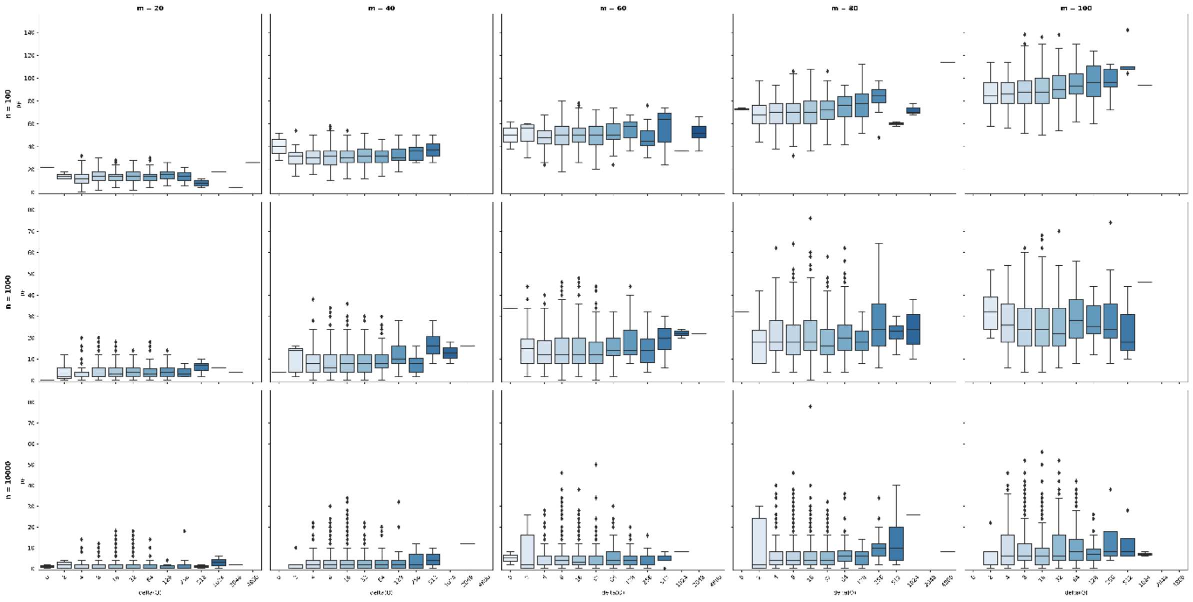


FIGURE A.8. The Robinson-Foulds metric as a function of the maximum Euclidian distance between the two matrices for each number of taxa (m) and number of site (n).

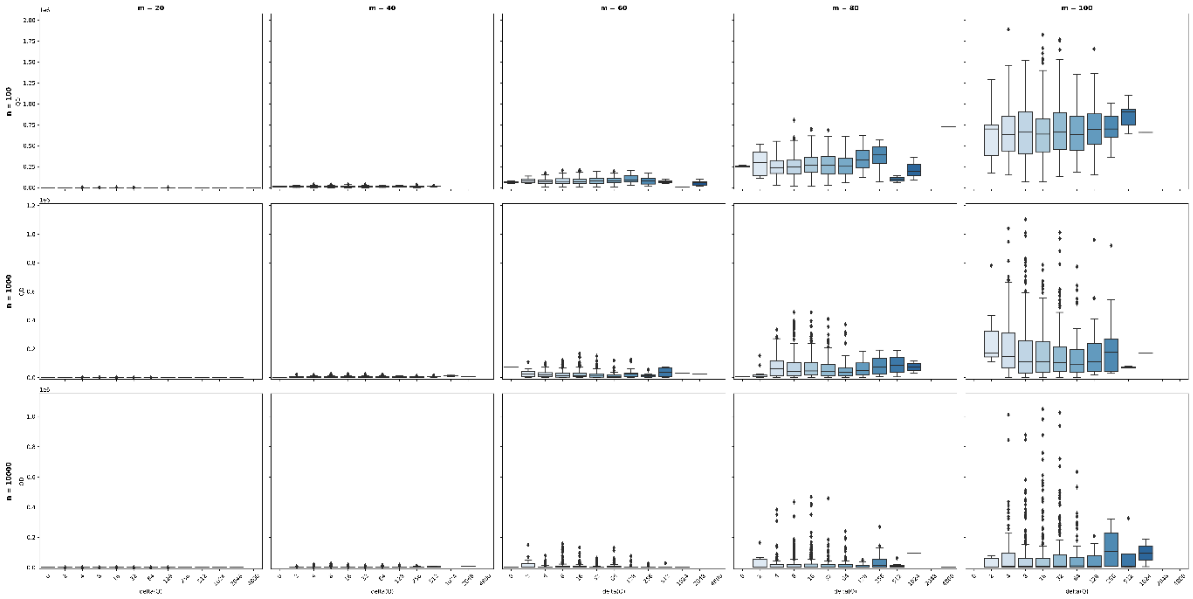


FIGURE A.9. The Quartet distance as a function of the maximum Euclidian distance between the two matrices for each number of taxa (m) and number of site (n).

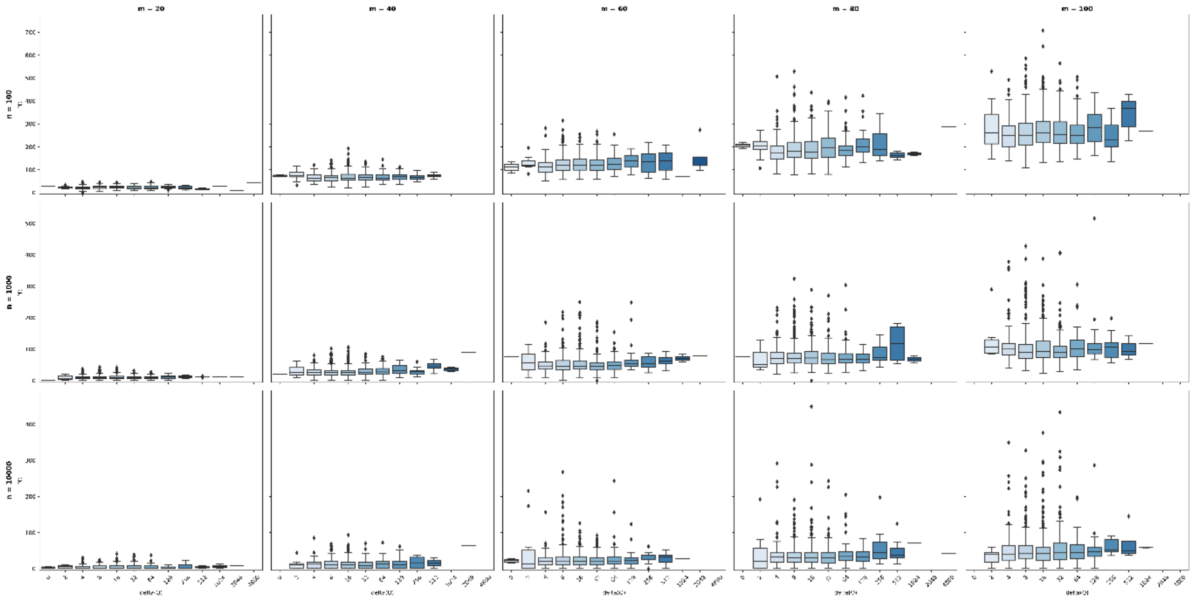


FIGURE A.10. The Path-Difference distance as a function of the maximum Euclidian distance between the two matrices for each number of taxa (m) and number of site (n).

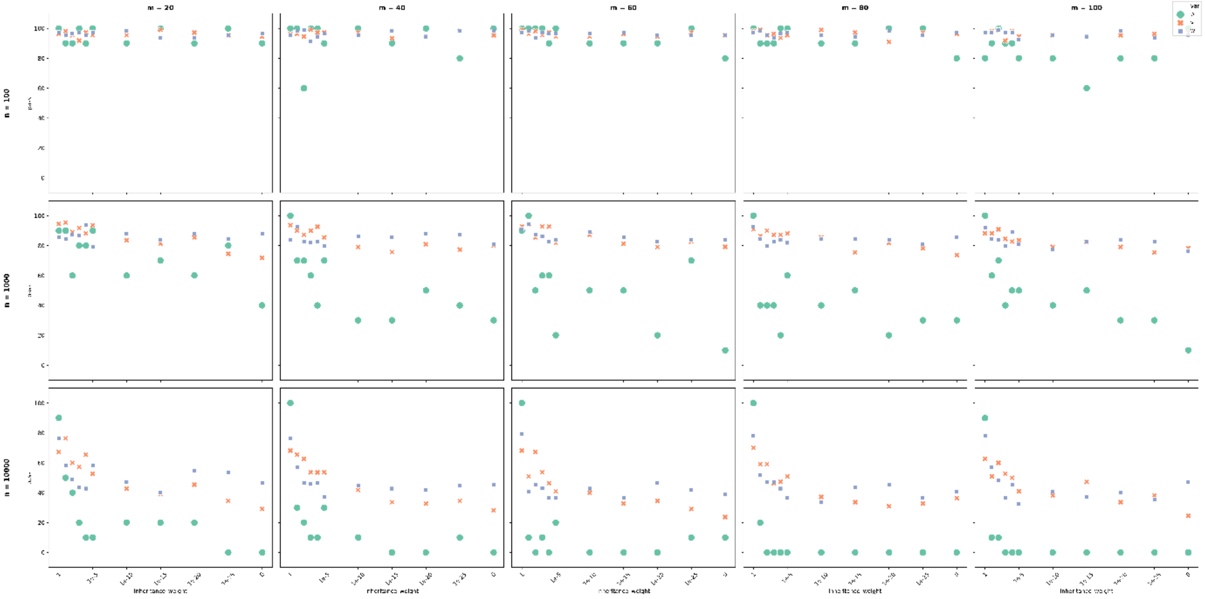


FIGURE A.11. The percentage of datasets that pass MaxSymTest as a function of the inheritance weight of the base frequencies (ν), the substitution model (ρ), the inheritance weight of the substitution rates (ω) for each number of taxa (m) and number of sites (n).

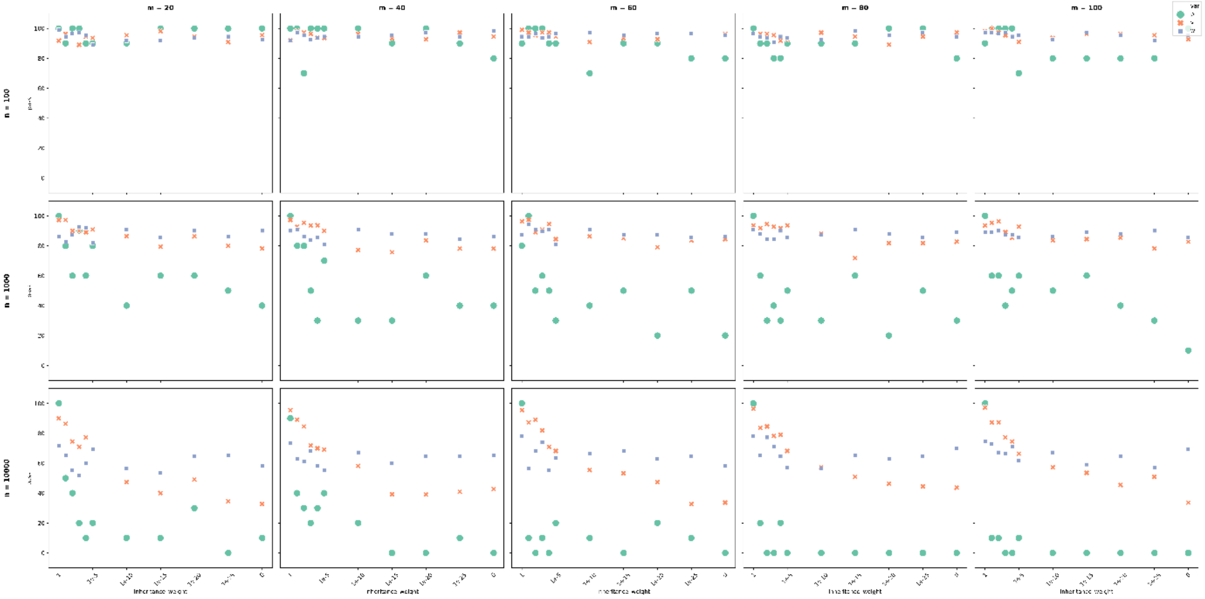


FIGURE A.12. The percentage of datasets that pass MaxSymTest$_{mar}$ as a function of the inheritance weight of the base frequencies (ν), the substitution model (ρ), the inheritance weight of the substitution rates (ω) for each number of taxa (m) and number of sites (n).

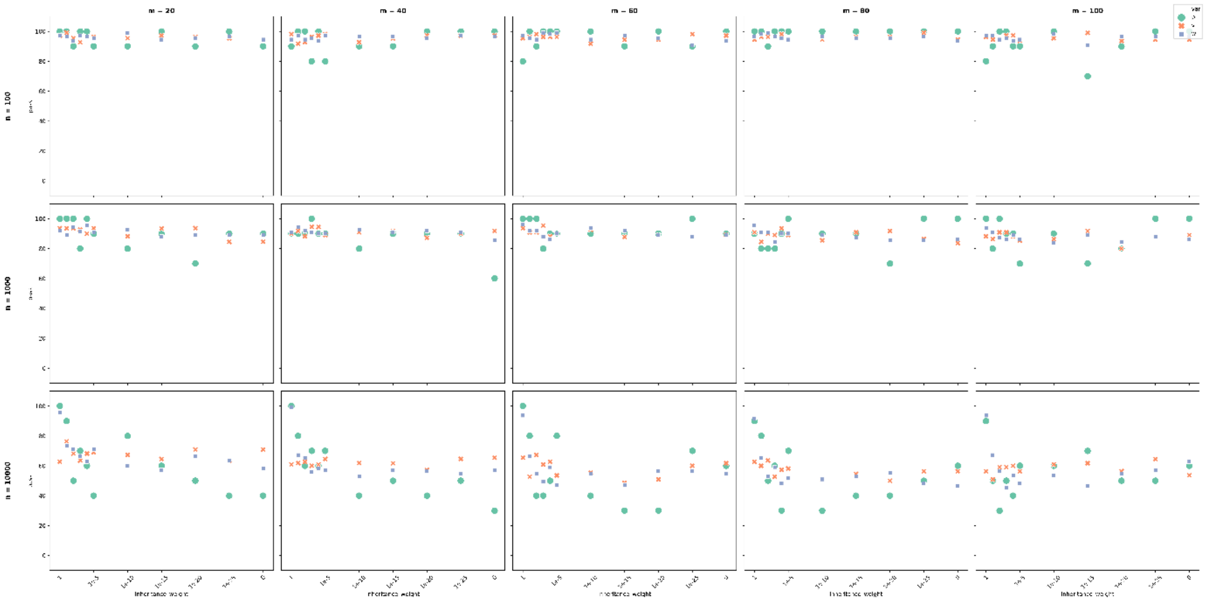


FIGURE A.13. The percentage of datasets that pass MaxSymTest$_{int}$ as a function of the inheritance weight of the base frequencies (ν), the substitution model (ρ), the inheritance weight of the substitution rates (ω) for each number of taxa (m) and number of sites (n).

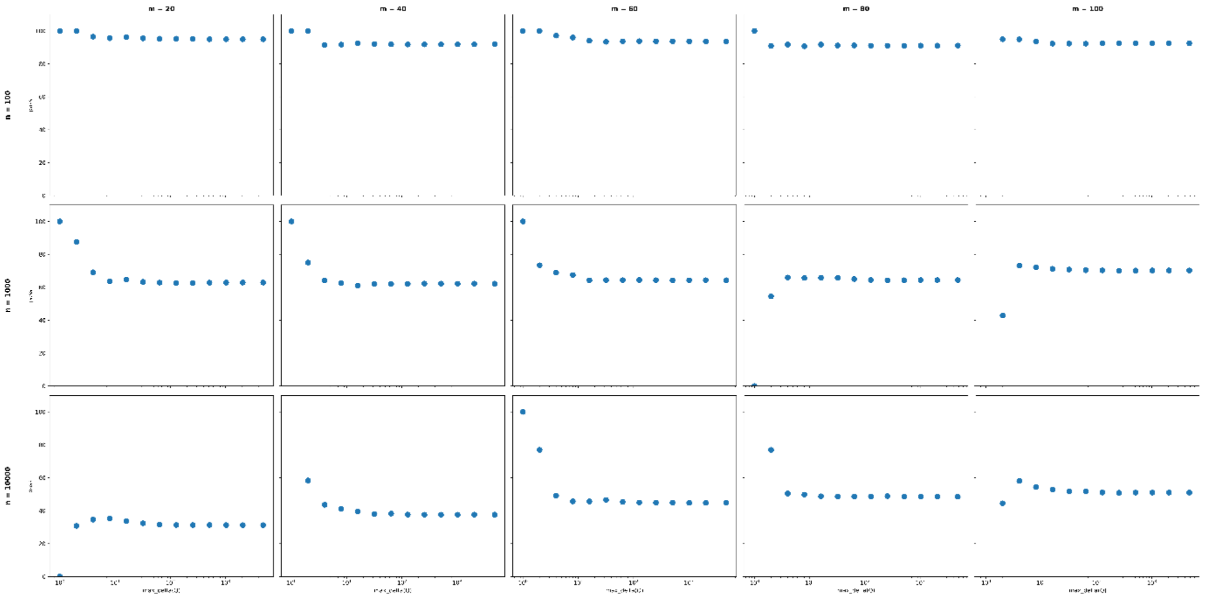


FIGURE A.14. The percentage of datasets that pass MaxSymTest as a function of the maximum Euclidian distance between the two matrices for each number of taxa (m) and number of site (n).

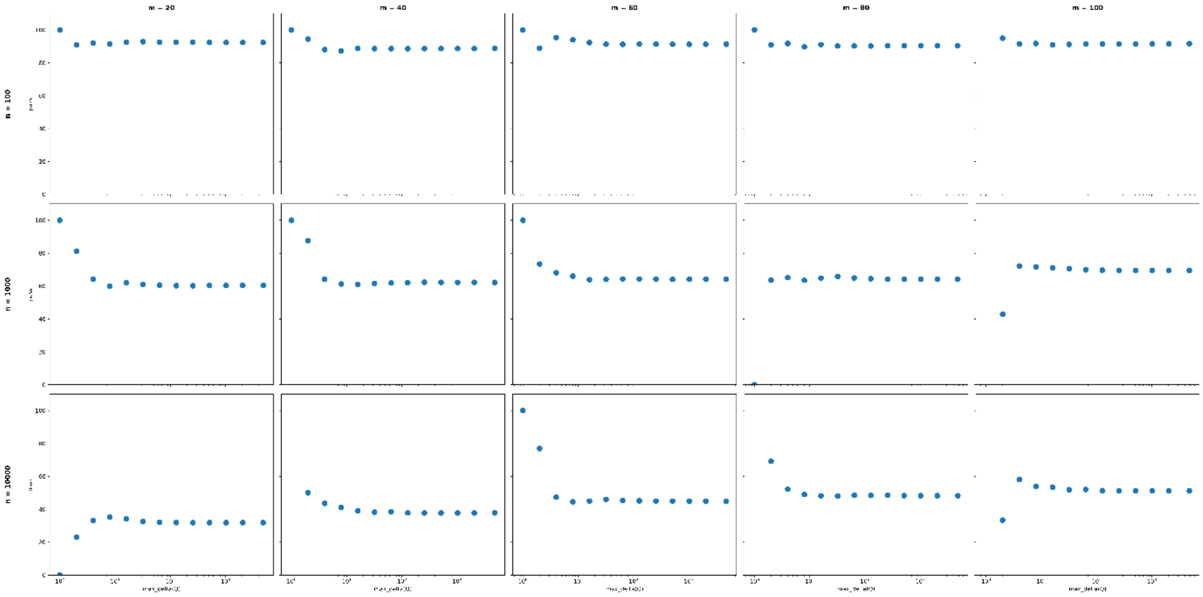


FIGURE A.15. The percentage of datasets that pass MaxSymTest$_{mar}$ as a function of the maximum Euclidian distance between the two matrices for each number of taxa (m) and number of site (n).

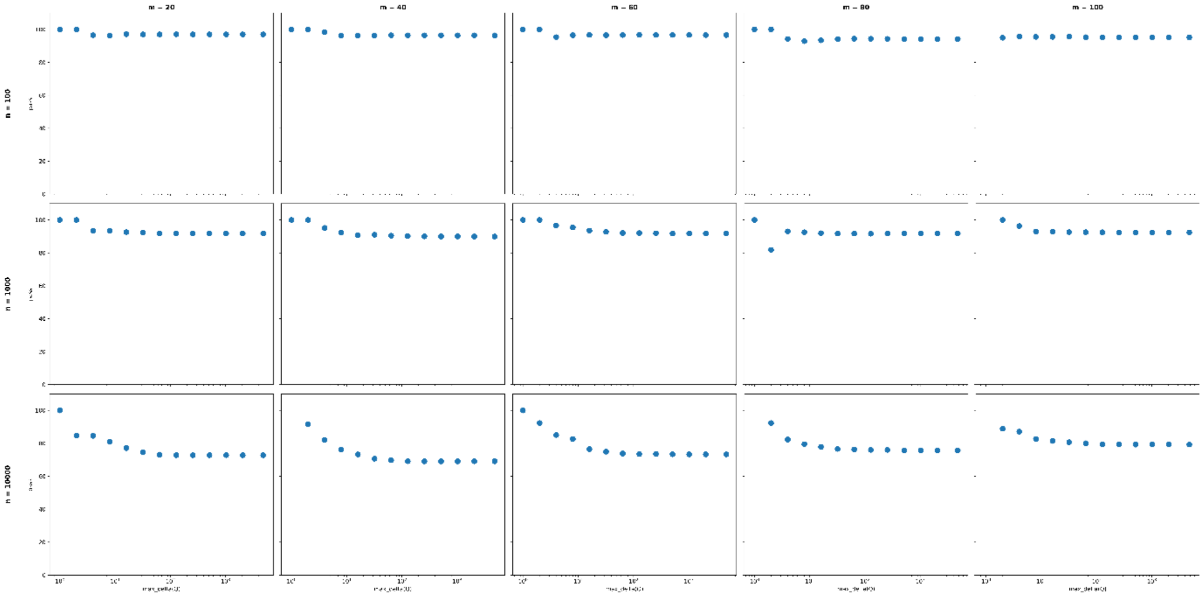


FIGURE A.16. The percentage of datasets that pass MaxSymTest$_{int}$ as a function of the maximum Euclidian distance between the two matrices for each number of taxa (m) and number of site (n).

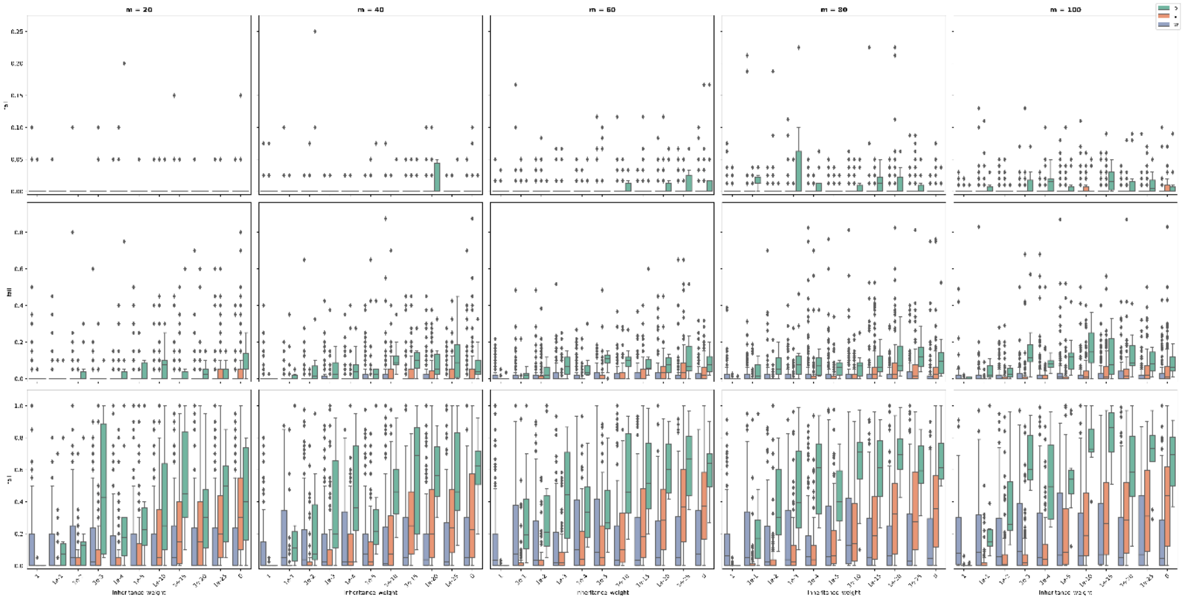


FIGURE A.17. The percentage of sequences that fail the Chi^2_{Rank} test in each dataset as a function of the inheritance weight of the base frequencies (ν), the substitution model (ρ), the inheritance weight of the substitution rates (ω) for each number of taxa (m) and number of sites (n).

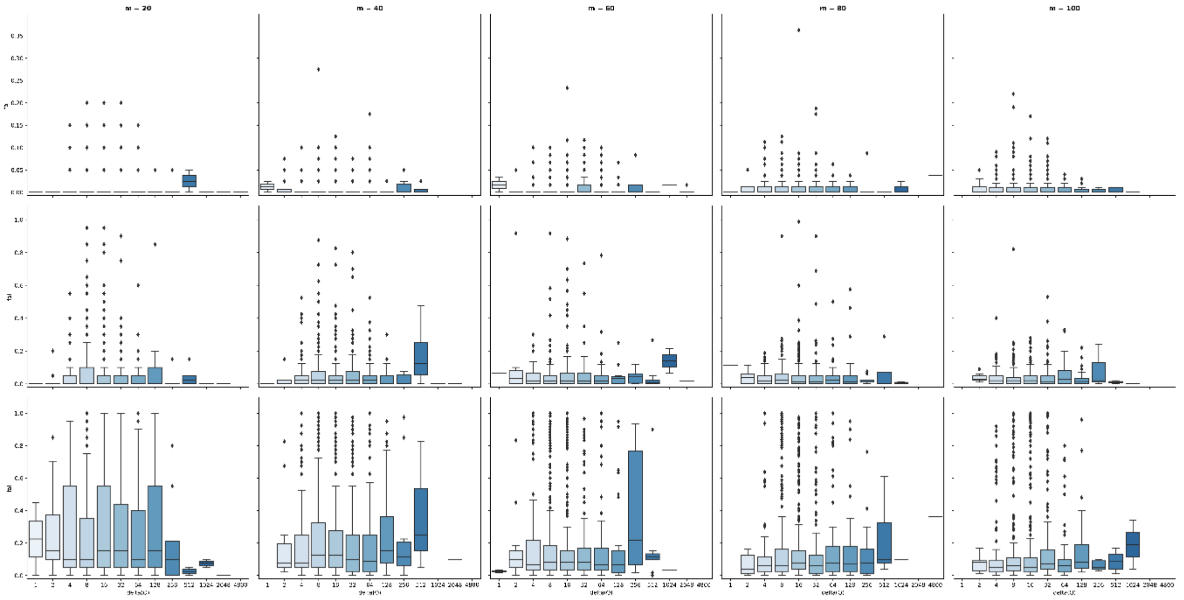


FIGURE A.18. The percentage of sequences that fail the Chi^2_{Rank} test in each dataset as a function of the Euclidian distance between the two matrices for each number of taxa (m) and number of site (n).

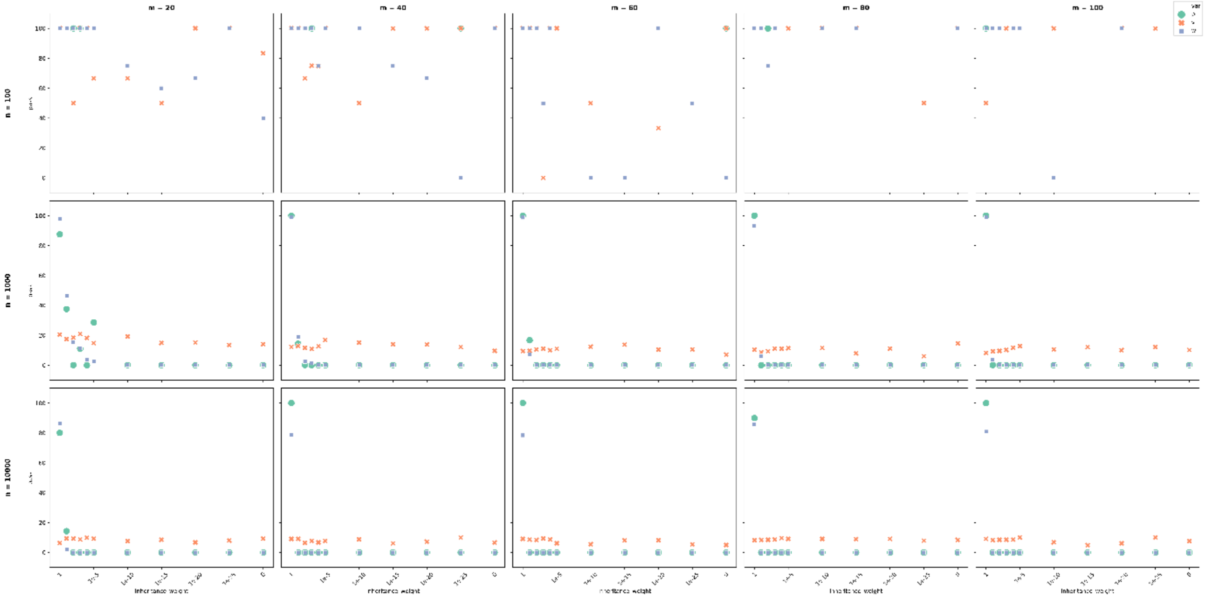


FIGURE A.19. The percentage of datasets that pass the WH test as a function of the inheritance weight of the base frequencies (ν), the substitution model (ρ), the inheritance weight of the substitution rates (ω) for each number of taxa (m) and number of sites (n).

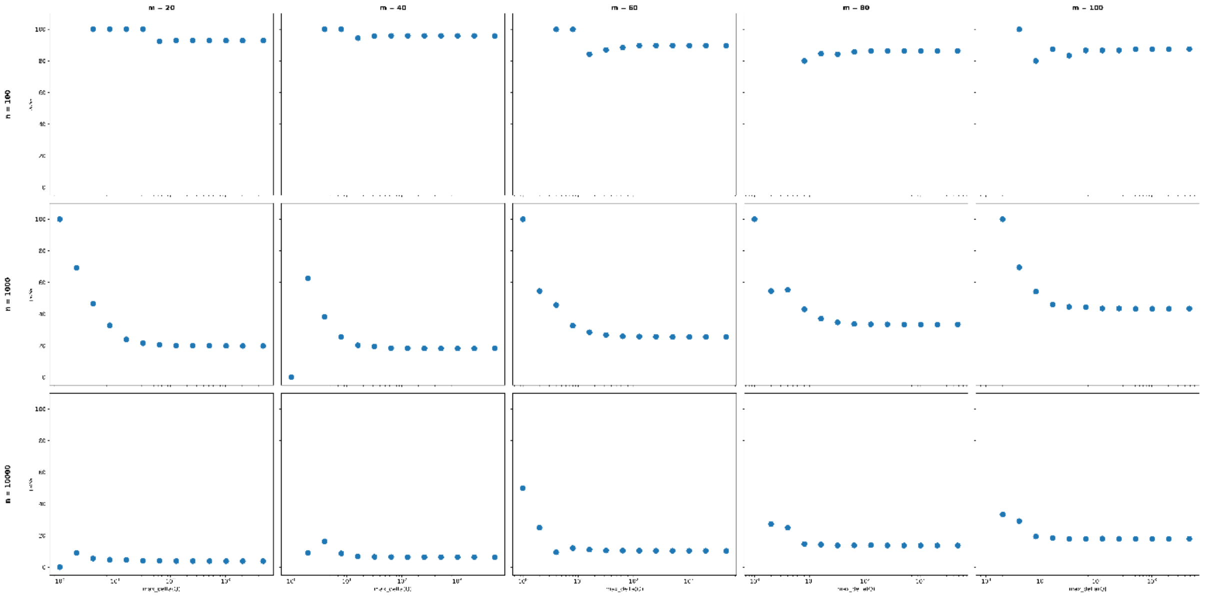


FIGURE A.20. The percentage of sequences that fail the WH test in each dataset as a function of the Euclidian distance between the two matrices for each number of taxa (m) and number of site (n).

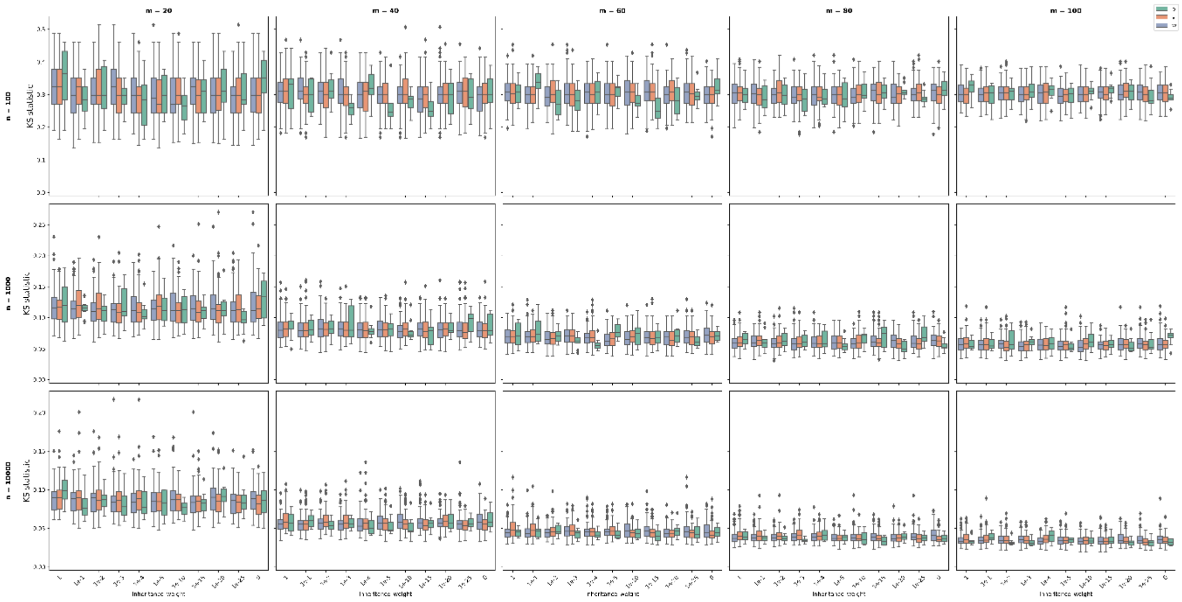


FIGURE A.21. Two-sample Kolmogorov-Smirnov test statistic as a function of the inheritance weight of the base frequencies (ν), the substitution model (ρ), the inheritance weight of the substitution rates (ω) for each number of taxa (m) and number of site (n).

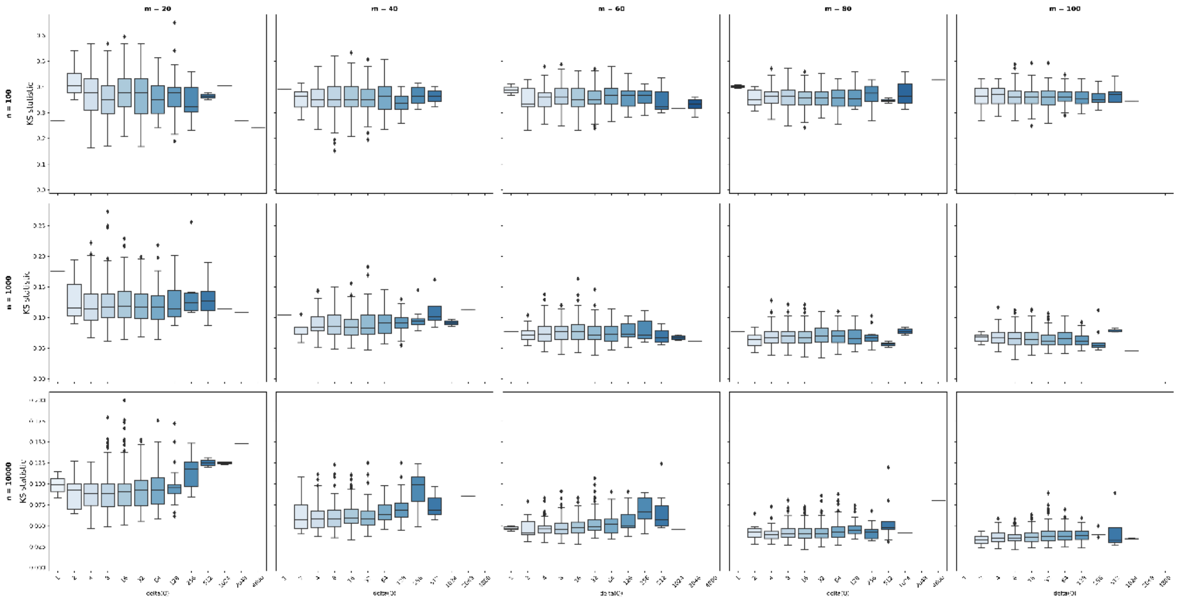


FIGURE A.22. Two-sample Kolmogorov-Smirnov test statistic as a function of the inheritance weight of the Euclidian distance between the two matrices for each number of taxa (m) and number of site (n).

References

Anderson FE, Bergman A, Cheng SH, Pankey MS, Valinassab T. 2013. Data from: Lights out: the evolution of bacterial bioluminescence in Loliginidae. In: Dryad Data Repository.

Anderson FE, Bergman A, Cheng SH, Pankey MS, Valinassab T. 2014. Lights out: the evolution of bacterial bioluminescence in Loliginidae. Hydrobiologia 725:189-203.

Ballesteros JA, Sharma PP. 2019a. A Critical Appraisal of the Placement of Xiphosura (Chelicerata) with Account of Known Sources of Phylogenetic Error. Syst. Biol.

Ballesteros JA, Sharma PP. 2019b. Data from: A critical appraisal of the placement of Xiphosura (Chelicerata) with account of known sources of phylogenetic error. In: Dryad.

Becker EA, Yao AI, Seitzer PM, Kind T, Wang T, Eigenheer R, Shao KS, Yarov-Yarovoy V, Facciotti MT. 2016. A Large and Phylogenetically Diverse Class of Type 1 Opsins Lacking a Canonical Retinal Binding Site. PLoS One 11:e0156543.

Becker EA, Yao AI, Seitzer PM, Kind T, Wang T, Eigenheer R, Shao KSY, Yarov-Yarovoy V, Facciotti MT. 2017. Data from: A large and phylogenetically diverse class of type 1 opsins lacking a canonical retinal binding site. In: Dryad.

Bergsten J, Nilsson AN, Ronquist F. 2013a. Bayesian tests of topology hypotheses with an example from diving beetles. Syst. Biol. 62:660-673.

Bergsten J, Nilsson AN, Ronquist F. 2013b. Data from: Bayesian tests of topology hypotheses with an example from diving beetles. In: Dryad Data Repository.

Borowiec ML, Lee EK, Chiu JC, Plachetzki DC. 2016. Data from: Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. In: Dryad Digital Repository.

Borowiec ML, Lee EK, Chiu JC, Plachetzki DC. 2015. Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. BMC Genomics 16:987.

Branstetter MG, Danforth BN, Pitts JP, Faircloth BC, Ward PS, Buffington ML, Gates MW, Kula RR, Brady SG. 2017a. Data from: Phylogenomic insights into the evolution of stinging wasps and the origins of ants and bees. In: Dryad Digital Repository.

Branstetter MG, Danforth BN, Pitts JP, Faircloth BC, Ward PS, Buffington ML, Gates MW, Kula RR, Brady SG. 2017b. Phylogenomic Insights into the Evolution of Stinging Wasps and the Origins of Ants and Bees. Curr. Biol. 27:1019-1025.

Broughton RE, Betancur RR, Li C, Arratia G, Orti G. 2013a. Data from: Multi-locus phylogenetic analysis reveals the pattern and tempo of bony fish evolution. In: Dryad Data Repository.

Broughton RE, Betancur RR, Li C, Arratia G, Orti G. 2013b. Multi-locus phylogenetic analysis reveals the pattern and tempo of bony fish evolution. PLoS Curr 5.

Brown RM, Siler CD, Das I, Min PY. 2012a. Data from: Testing the phylogenetic affinities of Southeast Asia’s rarest geckos: Flap-legged geckos (Luperosaurus), Flying geckos (Ptychozoon) and their relationship to the pan-Asian genus Gekko. In: Dryad Data Repository.

Brown RM, Siler CD, Das I, Min Y. 2012b. Testing the phylogenetic affinities of Southeast Asia's rarest geckos: Flap-legged geckos (Luperosaurus), Flying geckos (Ptychozoon) and their relationship to the pan-Asian genus Gekko. Mol Phylogenet Evol 63:915-921.

Cannon JT, Vellutini BC, Smith J, 3rd, Ronquist F, Jondelius U, Hejnol A. 2016a. Xenacoelomorpha is the sister group to Nephrozoa. Nature 530:89-93.

Cannon JT, Vellutini BC, Smith J, Ronquist F, Jondelius U, Hejnol A. 2016b. Data from: Xenacoelomorpha is the sister group to Nephrozoa. In: Dryad Data Repository.

Chen M-Y, Liang D, Zhang P. 2015a. Data from: Selecting question-specific genes to reduce incongruence in phylogenomics: a case study of jawed vertebrate backbone phylogeny. In: Dryad.

Chen MY, Liang D, Zhang P. 2015b. Selecting Question-Specific Genes to Reduce Incongruence in Phylogenomics: A Case Study of Jawed Vertebrate Backbone Phylogeny. Syst. Biol. 64:1104-1120.

Cognato AI, Vogler AP. 2001a. Data from: Exploring data interaction and nucleotide alignment in a multiple gene analysis of Ips (Coleoptera: Scolytinae). In: Dryad Data Repository.

Cognato AI, Vogler AP. 2001b. Exploring data interaction and nucleotide alignment in a multiple gene analysis of Ips (Coleoptera: Scolytinae). Syst. Biol. 50:758-780.

Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC. 2012a. Data from: More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. In: Dryad Digital Repository.

Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC. 2012b. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biol. Lett. 8:783-786.

Day JJ, Peart CR, Brown KJ, Friel JP, Bills R, Moritz T. 2013a. Continental diversification of an African catfish radiation (Mochokidae: Synodontis). Syst. Biol. 62:351-365.

Day JJ, Peart CR, Brown KJ, Friel JP, Bills R, Moritz T. 2013b. Data from: Continental diversification of an African catfish radiation (Mochokidae: Synodontis). In: Dryad Data Repository.

Devitt TJ, Cameron Devitt SE, Hollingsworth BD, McGuire JA, Moritz C. 2013. Data from: Montane refugia predict population genetic structure in the Large-blotched Ensatina salamander. In: Dryad Data Repository.

Devitt TJ, Devitt SE, Hollingsworth BD, McGuire JA, Moritz C. 2013. Montane refugia predict population genetic structure in the Large-blotched Ensatina salamander. Mol. Ecol. 22:1650-1665.

Dornburg A, Moore JA, Webster R, Warren DL, Brandley MC, Iglesias TL, Wainwright PC, Near TJ. 2012a. Data from: Molecular phylogenetics of squirrelfishes and soldierfishes (Teleostei:Beryciformes: Holocentridae): reconciling more than 100 years of taxonomic confusion. In: Dryad Data Repository.

Dornburg A, Moore JA, Webster R, Warren DL, Brandley MC, Iglesias TL, Wainwright PC, Near TJ. 2012b. Molecular phylogenetics of squirrelfishes and soldierfishes (Teleostei: Beryciformes: Holocentridae): reconciling more than 100 years of taxonomic confusion. Mol Phylogenet Evol 65:727-738.

Faircloth BC, Sorenson L, Santini F, Alfaro ME. 2013a. Data from: A phylogenomic perspective on the radiation of ray-finned fishes based upon targeted sequencing of ultraconserved elements (UCEs). In: Dryad Data Repository.

Faircloth BC, Sorenson L, Santini F, Alfaro ME. 2013b. A Phylogenomic Perspective on the Radiation of Ray-Finned Fishes Based upon Targeted Sequencing of Ultraconserved Elements (UCEs). PLoS One 8:e65923.

Fong JJ, Brown JM, Fujita MK, Boussau B. 2012a. Data from: A phylogenomic approach to vertebrate phylogeny supports a turtle-archosaur affinity and a possible paraphyletic Lissamphibia. In: Dryad Data Repository.

Fong JJ, Brown JM, Fujita MK, Boussau B. 2012b. A phylogenomic approach to vertebrate phylogeny supports a turtle-archosaur affinity and a possible paraphyletic lissamphibia. PLoS One 7:e48990.

Horn JW, Xi Z, Riina R, Peirson JA, Yang Y, Dorsey BL, Berry PE, Davis CC, Wurdack KJ. 2014a. Data from: Evolutionary bursts in Euphorbia (Euphorbiaceae) are linked with photosynthetic pathway. In: Dryad Data Repository.

Horn JW, Xi Z, Riina R, Peirson JA, Yang Y, Dorsey BL, Berry PE, Davis CC, Wurdack KJ. 2014b. Evolutionary bursts in Euphorbia (Euphorbiaceae) are linked with photosynthetic pathway. Evolution 68:3485-3504.

Irisarri I, Baurain D, Brinkmann H, Delsuc F, Sire J, Kupfer A, Petersen J, Jarek M, Meyer A, Vences M, et al. 2017a. Data from: Phylotranscriptomic consolidation of the jawed vertebrate timetree. In: Dryad Digital Repository.

Irisarri I, Baurain D, Brinkmann H, Delsuc F, Sire JY, Kupfer A, Petersen J, Jarek M, Meyer A, Vences M, et al. 2017b. Phylotranscriptomic consolidation of the jawed vertebrate timetree. Nat Ecol Evol 1:1370-1378.

Jarvis ED, Mirarab S, Aberer A, Houde P, Li C, Ho S, Faircloth BC, Nabholz B, Howard JT, Suh A, et al. 2014. Data from: Phylogenomic analyses data of the avian phylogenomics project. In: GigaScience Database.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT, et al. 2015. Phylogenomic analyses data of the avian phylogenomics project. Gigascience 4:4.

Kawahara AY, Rubinoff D. 2013a. Convergent evolution of morphology and habitat use in the explosive Hawaiian fancy case caterpillar radiation. J. Evol. Biol. 26:1763-1773.

Kawahara AY, Rubinoff D. 2013b. Data from: Convergent evolution in the explosive Hawaiian Fancy Cased caterpillar radiation. In: Dryad Data Repository.

Lartillot N, Delsuc F. 2012a. Data from: Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. In: Dryad Data Repository.

Lartillot N, Delsuc F. 2012b. Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. Evolution 66:1773-1787.

Leache AD, Chavez AS, Jones LN, Grummer JA, Gottscho AD, Linkem CW. 2015. Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. Genome Biol Evol 7:706-719.

Leaché AD, Chavez AS, Jones LN, Grummer JA, Gottscho AD, Linkem CW. 2015. Data from: Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. In: Dryad.

Looney BP, Ryberg M, Hampe F, Sanchez-Garcia M, Matheny PB. 2016. Into and out of the tropics: global diversification patterns in a hyperdiverse clade of ectomycorrhizal fungi. Mol. Ecol. 25:630-647.

Looney BP, Ryberg M, Hampe F, Sánchez-García M, Matheny PB. 2015. Data from: Into and out of the tropics: global diversification patterns in a hyper-diverse clade of ectomycorrhizal fungi. In: Dryad.

McCormack JE, Harvey MG, Faircloth BC, Crawford NG, Glenn TC, Brumfield RT. 2013a. Data from: A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. In: Dryad Data Repository.

McCormack JE, Harvey MG, Faircloth BC, Crawford NG, Glenn TC, Brumfield RT. 2013b. A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. PLoS One 8:e54848.

Meiklejohn KA, Faircloth BC, Glenn TC, Kimball RT, Braun EL. 2016a. Analysis of a Rapid Evolutionary Radiation Using Ultraconserved Elements: Evidence for a Bias in Some Multispecies Coalescent Methods. Syst. Biol. 65:612-627.

Meiklejohn KA, Faircloth BC, Glenn TC, Kimball RT, Braun EL. 2016b. Data from: Analysis of a rapid evolutionary radiation using ultraconserved elements (UCEs): Evidence for a bias in some multi-species coalescent methods. In: Dryad.

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, et al. 2014a. Data from: Phylogenomics resolves the timing and pattern of insect evolution. In: Dryad Digital Repository.

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, et al. 2014b. Phylogenomics resolves the timing and pattern of insect evolution. Science 346:763-767.

Moyle RG, Oliveros CH, Andersen MJ, Hosner PA, Benz BW, Manthey JD, Travers SL, Brown RM, Faircloth BC. 2016a. Data from: Tectonic collision and uplift of Wallacea triggered the global songbird radiation. In: Dryad Data Repository.

Moyle RG, Oliveros CH, Andersen MJ, Hosner PA, Benz BW, Manthey JD, Travers SL, Brown RM, Faircloth BC. 2016b. Tectonic collision and uplift of Wallacea triggered the global songbird radiation. Nat Commun 7:12709.

Murray EA, Carmichael AE, Heraty JM. 2013a. Ancient host shifts followed by host conservatism in a group of ant parasitoids. Proc Biol Sci 280:20130495.

Murray EA, Carmichael AE, Heraty JM. 2013b. Data from: Ancient host shifts followed by host conservatism in a group of ant parasitoids. In: Dryad Data Repository.

Near TJ, Dornburg A, Eytan RI, Keck BP, Smith WL, Kuhn KL, Moore JA, Price SA, Burbrink FT, Friedman M, et al. 2013a. Data from: Phylogeny and tempo of diversification in the superradiation of spiny-rayed fishes. In: Dryad.

Near TJ, Dornburg A, Eytan RI, Keck BP, Smith WL, Kuhn KL, Moore JA, Price SA, Burbrink FT, Friedman M, et al. 2013b. Phylogeny and tempo of diversification in the superradiation of spiny-rayed fishes. Proc Natl Acad Sci U S A 110:12738-12743.

Nguyen AD, Gotelli NJ, Cahan SH. 2016a. Data from: The evolution of heat shock protein sequences, cis-regulatory elements, and expression profiles in the eusocial Hymenoptera. In: Dryad.

Nguyen AD, Gotelli NJ, Cahan SH. 2016b. The evolution of heat shock protein sequences, cis-regulatory elements, and expression profiles in the eusocial Hymenoptera. BMC Evol. Biol. 16:15.

Oaks JR. 2011a. Data from: A time-calibrated species tree of Crocodylia reveals a recent radiation of the true crocodiles. In: Dryad Data Repository.

Oaks JR. 2011b. A time-calibrated species tree of Crocodylia reveals a recent radiation of the true crocodiles. Evolution 65:3285-3297.

Prebus M. 2017a. Data from: Insights into the evolution, biogeography and natural history of the acorn ants, genus Temnothorax Mayr (Hymenoptera: Formicidae). In: Dryad.

Prebus M. 2017b. Insights into the evolution, biogeography and natural history of the acorn ants, genus Temnothorax Mayr (hymenoptera: Formicidae). BMC Evol. Biol. 17:250.

Pyron RA, Wiens JJ. 2011. A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. Mol Phylogenet Evol 61:543-583.

Pyron RA, Wiens JJ, Alexander Pyron R. 2011. Data from: A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. In: Dryad.

Ran JH, Shen TT, Wang MM, Wang XQ. 2018a. Data from: Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between Gnetales and angiosperms. In: Dryad Digital Repository.

Ran JH, Shen TT, Wang MM, Wang XQ. 2018b. Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between Gnetales and angiosperms. Proc Biol Sci 285:20181012.

Reddy S, Kimball RT, Pandey A, Hosner PA, Braun MJ, Hackett SJ, Han K, Harshman J, Huddleston CJ, Kingston S, et al. 2017a. Data from: Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. In: Dryad Digital Repository.

Reddy S, Kimball RT, Pandey A, Hosner PA, Braun MJ, Hackett SJ, Han KL, Harshman J, Huddleston CJ, Kingston S, et al. 2017b. Why Do Phylogenomic Data Sets Yield Conflicting Trees? Data Type Influences the Avian Tree of Life more than Taxon Sampling. Syst. Biol. 66:857-879.

Richart CH, Hayashi CY, Hedin M. 2016a. Data from: Phylogenomic analyses resolve an ancient trichotomy at the base of Ischyropsalidoidea (Arachnida, Opiliones) despite high levels of gene tree conflict and unequal minority resolution frequencies. In: Dryad.

Richart CH, Hayashi CY, Hedin M. 2016b. Phylogenomic analyses resolve an ancient trichotomy at the base of Ischyropsalidoidea (Arachnida, Opiliones) despite high levels of gene tree conflict and unequal minority resolution frequencies. Mol Phylogenet Evol 95:171-182.

Rightmyer MG, Griswold T, Brady SG. 2013a. Data from: Phylogeny and systematics of the bee genus Osmia (Hymenoptera: Megachilidae) with emphasis on North American Melanosmia: subgenera, synonymies, and nesting biology revisited. In: Dryad Data Repository.

Rightmyer MG, Griswold T, Brady SG. 2013b. Phylogeny and systematics of the bee genus Osmia (Hymenoptera: Megachilidae) with emphasis on North American Melanosmia: subgenera, synonymies and nesting biology revisited. Syst. Entomol. 38:561-576.

Sauquet H, Ho SY, Gandolfo MA, Jordan GJ, Wilf P, Cantrill DJ, Bayly MJ, Bromham L, Brown GK, Carpenter RJ, 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.

Sauquet H, Ho SYW, Gandolfo MA, Jordan GJ, Wilf P, Cantrill DJ, Bayly MJ, Bromham L, Brown GK, Carpenter RJ, et al. 2011. Data from: Testing the impact of calibration on molecular divergence times using a fossil-rich group: the case of Nothofagus (Fagales). In: Dryad Data Repository.

Seago AE, Giorgi JA, Li J, Ślipiński A. 2011a. Data from: Phylogeny, classification and evolution of ladybird beetles (Coleoptera: Coccinellidae) based on simultaneous analysis of molecular and morphological data. In: Dryad Data Repository.

Seago AE, Giorgi JA, Li J, Ślipiński A. 2011b. Phylogeny, classification and evolution of ladybird beetles (Coleoptera: Coccinellidae) based on simultaneous analysis of molecular and morphological data. Mol. Phylogen. Evol. 60:137-151.

Sharanowski BJ, Dowling APG, Sharkey MJ. 2011a. Data from: Molecular phylogenetics of Braconidae (Hymenoptera: Ichneumonoidea) based on multiple nuclear genes and implications for classification. In: Dryad Data Repository.

Sharanowski BJ, Dowling APG, Sharkey MJ. 2011b. Molecular phylogenetics of Braconidae (Hymenoptera: Ichneumonoidea), based on multiple nuclear genes, and implications for classification. Syst. Entomol. 36:549-572.

Shen X-X. 2018. Data from: Tempo and mode of genome evolution in the budding yeast subphylum. In: Figshare.

Shen XX, Opulente DA, Kominek J, Zhou X, Steenwyk JL, Buh KV, Haase MAB, Wisecaver JH, Wang M, Doering DT, et al. 2018. Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. Cell 175:1533-1545 e1520.

Siler C, Brown RM, Oliveros CH, Santanen A. 2013. Data from: Multilocus phylogeny reveals unexpected diversification patterns in Asian Wolf Snakes (genus Lycodon). In: Dryad Data Repository.

Siler CD, Oliveros CH, Santanen A, Brown RM. 2013. Multilocus phylogeny reveals unexpected diversification patterns in Asian wolf snakes (genus Lycodon). Zool. Scr. 42:262-277.

Smith BT, Harvey MG, Faircloth BC, Glenn TC, Brumfield RT. 2014a. Data from: Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. In: Dryad Digital Repository.

Smith BT, Harvey MG, Faircloth BC, Glenn TC, Brumfield RT. 2014b. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. Syst. Biol. 63:83-95.

Tolley KA, Townsend TM, Vences M. 2013a. Data from: Large-scale phylogeny of chameleons suggests African origins and Eocene diversification. In: Dryad Data Repository.

Tolley KA, Townsend TM, Vences M. 2013b. Large-scale phylogeny of chameleons suggests African origins and Eocene diversification. Proc Biol Sci 280:20130184.

Unmack PJ, Allen GR, Johnson JB. 2013a. Data from: Phylogeny and biogeography of rainbowfishes (Melanotaeniidae) from Australia and New Guinea. In: Dryad Data Repository.

Unmack PJ, Allen GR, Johnson JB. 2013b. Phylogeny and biogeography of rainbowfishes (Melanotaeniidae) from Australia and New Guinea. Mol Phylogenet Evol 67:15-27.

Varga T, Krizsán K, Földi C, Dima B, Sánchez-García M, Sánchez-Ramírez S, Szöllősi GJ, Szarkándi JG, Papp V, Albert L, et al. 2019a. Data from: Megaphylogeny resolves global patterns of mushroom evolution. In: Dryad.

Varga T, Krizsán K, Földi C, Dima B, Sánchez-García M, Sánchez-Ramírez S, Szöllősi GJ, Szarkándi JG, Papp V, Albert L, et al. 2019b. Megaphylogeny resolves global patterns of mushroom evolution. Nat Ecol Evol 3:668-678.

Wainwright PC, Smith WL, Price SA, Tang KL, Sparks JS, Ferry LA, Kuhn KL, Near TJ. 2012b. Data from: The evolution of pharyngognathy: a phylogenetic and functional appraisal of the pharyngeal jaw key innovation in labroid fishes and beyond. In: Dryad Data Repository.

Wainwright PC, Smith WL, Price SA, Tang KL, Sparks JS, Ferry LA, Kuhn KL, Near TJ, Eytan RI. 2012a. The evolution of pharyngognathy: a phylogenetic and functional appraisal of the pharyngeal jaw key innovation in labroid fishes and beyond. Syst. Biol. 61:1001-1027.

Whelan NV, Kocot KM, Moroz TP, Mukherjee K, Williams P, Paulay G, Moroz LL, Halanych KM. 2017a. Author Correction: Ctenophore relationships and their placement as the sister group to all other animals. Nat Ecol Evol 1:1783.

Whelan NV, Kocot KM, Moroz TP, Mukherjee K, Williams P, Paulay G, Moroz LL, Halanych KM. 2017b. Data from: Ctenophora Phylogeny Datasets and Core Orthologs. In: Figshare.

Wood HM, Matzke NJ, Gillespie RG, Griswold CE. 2012. Data from: Treating fossils as terminal taxa in divergence time estimation reveals ancient vicariance patterns in the palpimanoid spiders. In: Dryad Data Repository.

Wood HM, Matzke NJ, Gillespie RG, Griswold CE. 2013. Treating fossils as terminal taxa in divergence time estimation reveals ancient vicariance patterns in the palpimanoid spiders. Syst. Biol. 62:264-284.

Worobey M, Han G, Rambaut A. 2014a. Data from: A synchronized global sweep of the internal genes of modern avian influenza virus. In: Dryad Data Repository.

Worobey M, Han GZ, Rambaut A. 2014b. A synchronized global sweep of the internal genes of modern avian influenza virus. Nature 508:254-257.

Wu S, Edwards S, Liu L. 2019. Data from: Genome-scale DNA sequence data and the evolutionary history of placental mammals. In: Figshare.

Wu S, Edwards S, Liu L. 2018. Genome-scale DNA sequence data and the evolutionary history of placental mammals. Data Brief 18:1972-1975.

CHAPTER 3

HOW MANY PROCESSES ARE ENOUGH TO REPRESENT EVOLUTION?

Suha Naser-Khdour*[1], Bui Quang Minh[2], and Robert Lanfear[1]

(5)Department of Ecology and Evolution, Research School of Biology, Australian National University, Canberra, Australian Capital Territory, Australia

(6)School of Computing, Australian National University, Canberra, Australian Capital Territory, Australia

**Author for Correspondence: E-mail:* suha.naser@anu.edu.au

Contributions:

SNK wrote the python script, performed the analysis, analysed and interpreted the results, drafted the manuscript, and submitted the article for publication. MB contributed to the research design, conceptual development, and editorial comments. RL contributed to the research design, conceptual development and editorial comments.

Abstract

It is widely accepted that different genes and loci evolve under different substitution processes. Yet, most phylogenetic analyses still use homogeneous-time-reversible substitution models to infer evolutionary relationships. Several non-homogeneous and non-stationary models of nucleotide and amino acid evolution have been developed, but they are not widely popular mainly due to their high computational requirements. Here, we introduce a simple new user-friendly algorithm to find the best-fit non-homogeneous model in a Maximum Likelihood framework and apply that algorithm to three big empirical published datasets. Our results show that non-homogeneous models always outperform homogeneous models. In addition, we show that even a simple non-homogeneous model with only three matrices operating along the tree can significantly improve the goodness-of-fit in terms of AIC score. The algorithm is available in https://github.com/suhanaser/Non-Homogeneous-Model.

Keywords: phylogenetic inference, nonhomogeneous model, systematic bias, phylogeny

Introduction

It is widely accepted that most evolutionary processes operating along phylogenetic trees are neither stationary nor homogeneous. For example, simple distance-based approaches recently confirmed that this is the case across a wide variety of empirical phylogenetic datasets (Naser-Khdour, et al. 2019). Nonetheless, the ease of using stationary, reversible and homogeneous (SRH) models of evolution, and their robustness for topology inference in simulated datasets (Yang 2006; Naser-Khdour, et al. 2021), combined with their computational tractability, make them the most popular models in phylogenetic inference (Swofford 2001; Drummond and Rambaut 2007; Guindon, et al. 2010; Ronquist, et al. 2012; Bazinet, et al. 2014; Bouckaert, et al. 2014; Stamatakis 2014; Nguyen, et al. 2015; Höhna, et al. 2016).

The homogeneity assumption implies that the same evolutionary process operates along all the lineages in the phylogeny, in other words, it implies that one substitution matrix is used across the whole dataset. In order to relax this assumption, the evolutionary model should allow for different substitution processes to operate along the tree.

A growing body of evidence shows that most empirical datasets are not homogeneous. Several non-homogeneous substitution models have been introduced in the literature (Yang and Roberts 1995; Galtier and Gouy 1998; Galtier, et al. 1999; Foster 2004; Jayaswal, et al. 2005; Blanquart and Lartillot 2006; Jayaswal, et al. 2007; Blanquart and Lartillot 2008; Dutheil and Boussau 2008; Jayaswal, Jermiin, et al. 2011; Dutheil, et al. 2012; Zou, et al. 2012; Groussin, et al. 2013; Jayaswal, et al. 2014) but they are rarely used in empirical studies. The vast majority of these nonhomogeneous models focus on fitting different substitution matrices to different pre-defined groups of branches on a phylogenetic tree, but this approach suffers from requiring the user to specify appropriate groups of branches a-priori. On the other hand, some approaches use algorithms to find optimal groupings of branches; such as Bio++ (Dutheil and Boussau

2008), BppML (Groussin, et al. 2013), and HAL-HAS (Jayaswal, et al. 2014); these software are rarely used, mainly because of their relatively long run time. Yet, these approaches remain computationally intractable for large datasets. Due to these limitations, these models are rarely applied in empirical phylogenetic studies. Because of this, the extent and consequences of non-homogeneity in phylogenetic inference remain relatively unknown.

In this study, we use a greedy algorithm to examine how many Q matrices form the best fit for each dataset. Unlike other studies that focus on compositional heterogeneity across sites (e.g. Lartillot and Philippe 2004), our focus in this study is on compositional heterogeneity across lineages. Existing methods to ask this question (Dutheil and Boussau 2008; Groussin, et al. 2013; Jayaswal, et al. 2014), were not computationally feasible to run on datasets of the size used in this study. Thus, we instead designed a new approach that leverages the recently-released QMaker software. We introduce an approach to test for homogeneity among related clades based on the AIC score (Akaike 1974). This approach relies on first fitting independent standard stationary, reversible, and homogeneous (SRH) substitution models to pre-defined clades of taxa, and then using the AIC score to determine whether models with fewer Q matrices (where a single matrix is estimated from more than one pre-defined clade) are a better fit to the data. Applying this approach to our three empirical datasets reveals substantial evidence for non-homogeneity even among closely-related clades, but also shows that accounting for this non-homogeneity makes no appreciable difference to phylogenetic inference in this framework.

Material and Methods

Empirical dataset selection and clade definition

We used three publicly available empirical datasets of mammals, birds, and plants. These datasets were selected because of the high quality of their alignments, their size, and their taxonomic breadth. All of these factors contribute to the accuracy, power, and generality of any test of the homogeneity assumption in phylogenetics. For each dataset, we selected as many monophyletic sub-clades as possible, providing that the monophyly of each sub-clade has been well supported and non-controversial in previous studies, received 100% bootstrap support in our analyses, and contained at least 3 taxa (the minimum required to estimate a Q matrix; further details in Appendix Figs A.1-5).

a. The mammals' dataset (Wu, et al. 2018, 2019) comprises data from 82 species and over 3 million amino acid sites. For the purpose of this study, we considered 54 taxa that form 10 different clades: – Apes (6 taxa), Old World monkeys (4 taxa), Hystricomorpha (4 taxa), Myomorpha (7 taxa), Yinpterochiroptera (5 taxa), Yangochiroptera (4 taxa), Cetacea (5 taxa), Artiodactyla (5 taxa), Carnivora (7 taxa) and Afrotheria (7 taxa).

b. The birds' dataset (Jarvis, et al. 2014; Jarvis, et al. 2015) comprises data from 48 species and over 4 million amino acid sites. For the purpose of this study, we considered 32 taxa in 8 different clades – Galloanserae(3 taxa), Caprimulgiformes (3 taxa), Columbimorphae (3 taxa), Pelecaniformes (4 taxa), Procellariiformes (3 taxa), Coraciiformes (6 taxa), Passeriformes (7 taxa) and Accipitriformes (3 taxa).

c. The plants' dataset (Ran, et al. 2018b, a) comprises data from 35 species and over 400K amino acid sites. For the purpose of this study, we used all the 35 taxa grouped into 5 clades – Angiosperms (13 taxa), Cycadales (3 taxa), Conifer II (13 taxa), Gnetales (3 taxa) and Pinaceae (3 taxa).

For the mammals and the birds' datasets, we randomly subsample 10% of the sites to decrease the computational burden.

Defining clades for each dataset

In this study, we sought to test the null hypothesis that molecular evolution is homogeneous for each of the three datasets. That is, we ask whether the data are best fit by a single model of molecular evolution applied to all branches, as compared to having more than one model, with each applied to a subset of the branches in the tree.

To do this, we first made the necessary assumption that each of the pre-defined clades mentioned in the previous section would have at most one Q matrix applied to it. Then for each dataset, we set *"MaxQ"* to be the number of clades in that dataset.

Inferring the SRH model

For each dataset, we first estimated a fully homogeneous model which assumes that all of the branches represented in every sub-clade evolved under the same process. To do this, we used IQ-TREE 2 (Minh, et al. 2020) and ModelFinder (Kalyaanamoorthy, et al. 2017) with the best-fit fully partitioned model (Chernomor, et al. 2016) and edge-linked substitution rates (Duchene, et al. 2020). See commandline 1 in the Appendix.

We then used QMaker (Minh, et al. 2021) to estimate the empirical Q matrix and stationary frequencies for the homogeneous model using commandline 2 (Appendix).

Inferring the non-homogeneous model with MaxQ different matrices

For each dataset we then infer the best SRH model for each clade separately using commandline 3 (Appendix).

We then use QMaker (Minh, et al. 2021) to estimate the empirical Q matrix and stationary frequencies vector for each clade using commandline 4 (Appendix).

Next, we calculate the AIC score for the joint model that consists of all the clades' matrices according to equation 1:

$$1)\quad 2\sum_j k - 2\sum_j \ln(L)$$

Where k is the number of free parameters and L is the maximum value of the likelihood function for each clade j.

The greedy algorithm

Given the high computational costs of inferring Q matrices, it was computationally infeasible to evaluate all groupings of clades, and thus every possible combination of clades into fewer than MaxQ matrices. Therefore, in order to search for the best model in the space between the fully homogeneous model with one matrix and the fully non-homogeneous model with MaxQ matrices, we use a greedy algorithm. This algorithm starts with a model Q assigned to each clade (from the previous step) and iteratively merges the single pair of clades together which maximises the improvement in the AIC score. I.e. in the first step, we evaluate all the different combinations of two clades, then we merge the two clades which give the largest improvement in the AIC value. We repeat this merging process until we have one clade that represents all the taxa. Figure 1 shows an example of the process for MaxQ = 6 models. We continue merging clades even if the best merging makes the AIC score worse because we wish to evaluate the AIC score at every step of the algorithm.

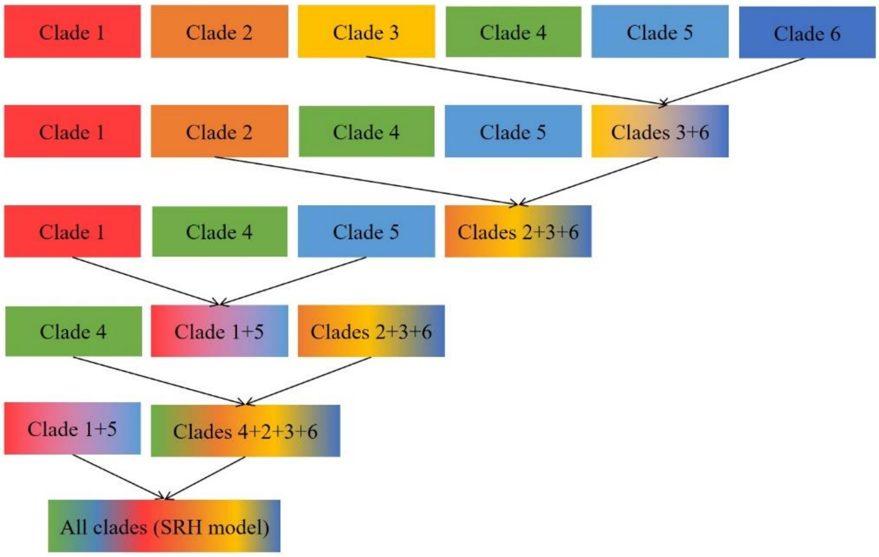


Figure 1| Schematic overview of the greedy algorithm with MaxQ = 6 matrices.

The algorithm:

1. Let *S* be the set of all clades in dataset *D*. For each clade in *S*, infer the best SRH model using command lines 3 and 4, and calculate $\text{AIC}_{\text{non-SRH}}$ for the joint model according to equation 1.
2. Run over all the possible combinations of two clades in *S* and calculate the AIC score for each pair.
3. Let *v* and *w* be the two merged clades from the previous step with the best AIC score ($\text{AIC}_{\text{merge}}$) into one single clade *M*. If $\text{AIC}_{\text{merge}} < \text{AIC}_{\text{non-SRH}}$, then update $\text{AIC}_{\text{non-SRH}} = \text{AIC}_{\text{merge}}$ and $S = S - v - w + M$. Otherwise, go to step 5.
4. If $S > 2$, go to step 2. Otherwise, go to step 5.
5. For dataset *D* infer the homogeneous model using command lines 1 and 2.

Clade clustering

Using a two-dimensional Principal Component Analysis (PCA) with all matrices from all datasets, we check the divergence within and between datasets. For each dataset, we perform a two-dimensional PCA with the homogeneous model, the relevant Q matrix (i.e. Q.Birds, Q.Mammals, Q.Plants) from IQ-TREE2 (Minh, et al. 2020) and all the clades' matrices. In addition, we performed another two-dimensional PCA with all the clades' matrices from all three datasets in order to check if clades from the same datasets cluster together or not.

Results

The fully homogeneous model always has a higher AICscore than non-homogeneous models

In the three datasets, the homogeneous model with one matrix for all clades has a notably higher AIC score than the other non-homogeneous model with at least two different matrices (Figure 2). In the mammals' dataset, the AIC value of the homogeneous model is more than 3600 scores higher than the best 2-matrices model where Afrotheria and Rodentia have one matrix and all the other clades have another matrix. In addition, the AIC value of the homogeneous model is more than 5600 scores higher than the best-fit model with 7 matrices. Yet, the differences between the best 5-matrices model, 6-matrices model, the best 8-matrices model, the best 9-matrices model and the best-fit model are small; 120, 27, 19, and 84 scores, respectively.

In the plants' dataset, the AIC value of the homogeneous model is more than 1600 scores higher than the best 2-matrices models where Angiosperms and Gnetales have one matrix and all other clades have a separate matrix. Furthermore, the AIC value of the homogeneous model is more than 18,300 scores higher than the best-fit model with 4 different matrices. In contrast to the mammals' dataset where the differences between the best-fit model and other close models are small, in the plants' dataset, these differences are much higher; the difference between the best-fit model and the best 3-matrices and the best 5-matrices models are 900 and 330, respectively.

In the birds' dataset, the AIC value of the homogeneous model is more than 1890 scores higher than the best 2-matrices models where Galloanserae, Caprimulgiformes and Passeriformes have one matrix and all other clades have another matrix. Moreover, the AIC value of the homogeneous model is more than 3500 scores higher than the best-fit model with 3 different matrices. Like the plants' dataset, in the bird's dataset, the differences between the best-fit model and the best 2-matrices and the best 4-matrices models are much higher than in the mammals' dataset and are 800 and 400, respectively.

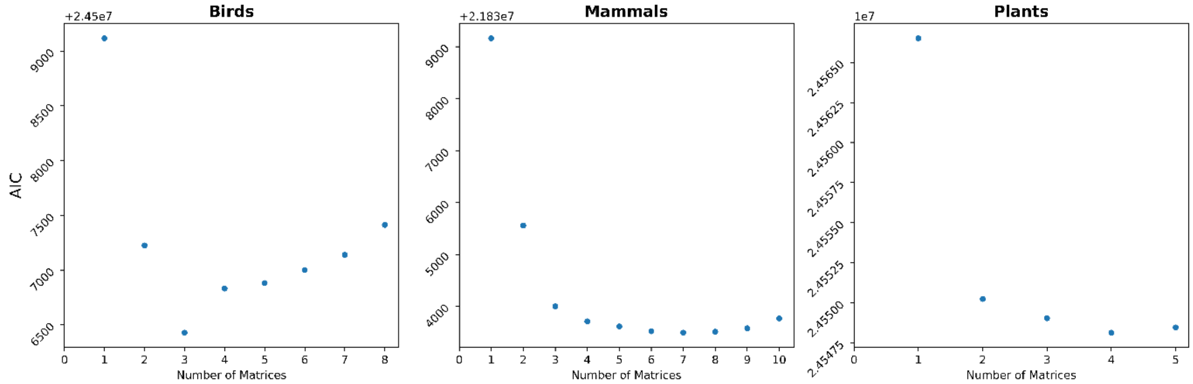


Figure 2| AIC score for each dataset as a function of the number of matrices in the model.

The fully non-homogeneous model is always worse than a simpler model

In all the datasets, there is at least one model that outperforms the fully non-homogeneous model with MinQ matrices. In the plants' datasets, the model with MaxQ – 1 matrices outperform the fully non-homogeneous model with MaxQ matrices. In the mammals' dataset, the model with MaxQ – 3 matrices was the best fit model and in the birds' dataset, the model with MaxQ – 5 matrices is the best model (Figure 2). The differences between the best-fit model and the fully non-homogeneous model with MaxQ matrices are 275, 327, and 985 for mammals, plants and birds, respectively. The differences are much higher when we compare the fully non-homogeneous model and the homogeneous model; 5400, 18,000 and 1700 for mammals, plants and birds, respectively.

Clades' matrices from the same dataset clusters together

A two-dimensional Principal Component Analysis (PCA) of all the matrices and the stationary frequencies vectors in the best-fit models shows defined clusters of clades, especially in terms of stationary frequencies (**Figure**). Yet, some datasets are more divergent than others. E.g. in terms of substitution rates, some of the clades in the mammals dataset are closer to clades from the birds and plants datasets than other mammalian clades.

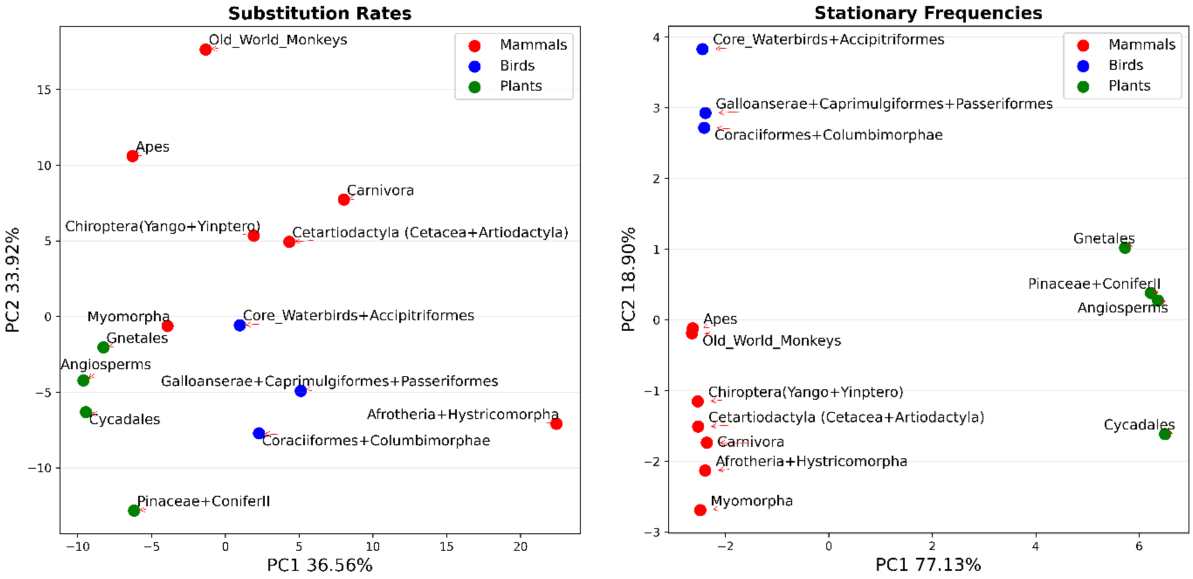


Figure 3| 2-D PCA for all the matrices and stationary frequencies vectors in the best model for each dataset.

Adding the matrix of all clades from the homogeneous model (i.e Q.Homogeneous) and the relevant Q matrix from IQ-TREE2 (i.e Q.Birds, Q.Mammals and Q.Plants) to the 2-D PCA we see a difference between the All matrix and the relevant Q Matrix (Figure 3).

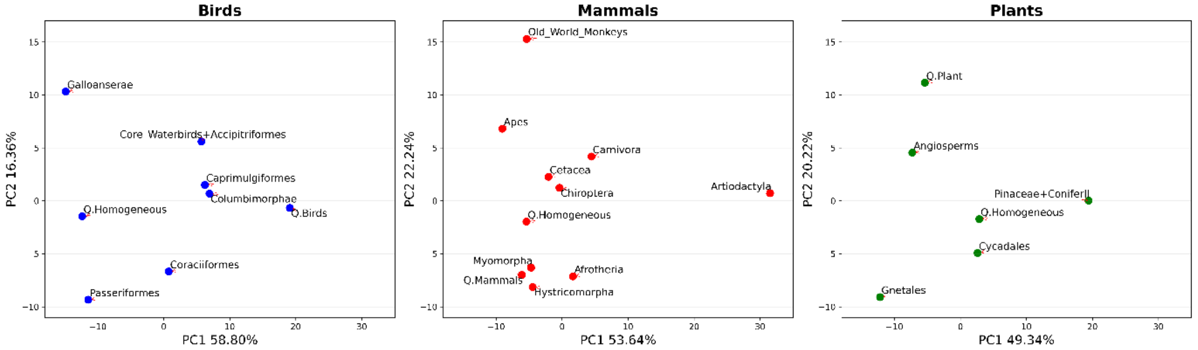


Figure 3| 2-D PCA for each dataset with the homogeneous model (substitution matrix + stationary frequencies vector).

Phylogenetic inference is unaffected by using different matrices

For each clade in each dataset, we inferred the phylogeny using the homogeneous model and the specific non-homogeneous model for that clade. Our results show that both phylogenies were identical; Robinson-Foulds distance is zero for all clades in all datasets. Moreover, using the Kolmogorov-Smirnov (KS) test we see that there is no significant difference between the two distributions of branch lengths (Figure 4).

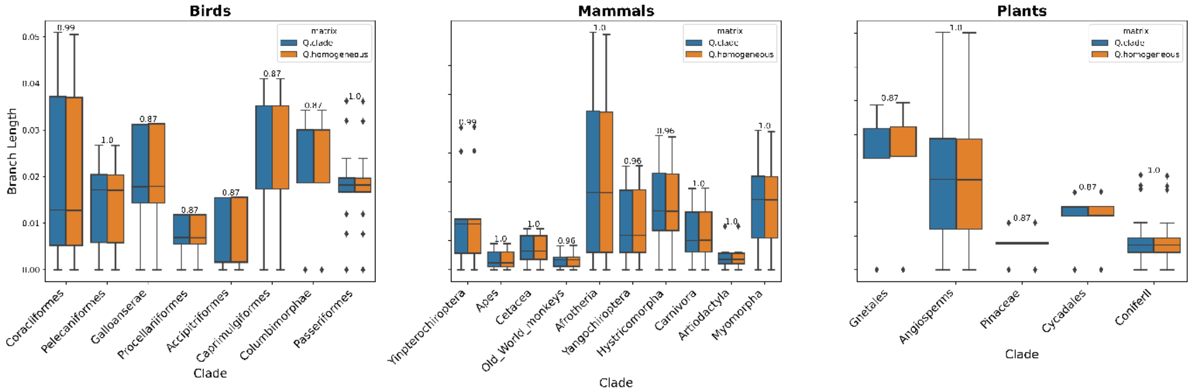


Figure 4| Branch lengths distribution of each phylogeny. The numbers on top of the boxes are the p-value from the KS test when we compare the two distributions of branch lengths.

Discussion

In this article, we tested for homogeneity of the molecular evolutionary processes among closely-related clades of birds, mammals, and plants using large multiple-sequence alignments. Our results suggest that while substantial and statistically significant heterogeneity exists, even between closely related clades, this heterogeneity has a little detectable effect on phylogenetic reconstructions of tree topologies or branch lengths.

Our results are consistent with previous studies that show that non-homogeneous models can capture the evolutionary signals better than homogeneous models. Using Maximum

Parsimony, Maximum Likelihood, Bayesian inference, and Neighbour-Joining methods to estimate candidates starting trees for the γ-Proteobacteria phylogeny, Herbeck et al. (Herbeck, et al. 2005) calculated the likelihood value for each tree under the non-homogeneous model of (Galtier and Gouy 1995). Their results show that the likelihood ratio of the non-homogeneous model was the best fit for all datasets. In addition, the results show that the phylogeny inferred under the ML method was very similar to the phylogeny inferred under the non-homogeneous model. Emphasizing our conclusion that ML inference methods can be very robust to non-homogeneous evolution, even when using SRH substitution models.

Similar to Herbeck et al., other studies that compared homogeneous and non-homogeneous models e.g. (Blanquart and Lartillot 2006; Boussau and Gouy 2006; Blanquart and Lartillot 2008; Boussau, et al. 2008; Dutheil and Boussau 2008; Jayaswal, Jermiin, et al. 2011; Zhang, et al. 2011; Dutheil, et al. 2012; Groussin, et al. 2013; Jayaswal, et al. 2014) showed a better fit of the non-homogeneous models over the homogeneous model. Looking at a two-dimensional PCA of all the matrices and the stationary frequencies vectors reveals that the differences between stationary frequencies are more pronounced between rather than within major clades, while the opposite seems to be true for the Q matrices. This conclusion agrees with previous studies that showed that non-stationarity is the major cause of systematic bias due to non-SRH evolution in phylogenetics (Galtier and Gouy 1998; Galtier, et al. 1999; Foster 2004; Jermiin, et al. 2004; Jayaswal, et al. 2005; Ababneh, et al. 2006; Blanquart and Lartillot 2008; Boussau, et al. 2008; Song, et al. 2010; Naser-Khdour, et al. 2019). Our results add to these, and demonstrate in addition that the Q matrices can differ substantially, even between closely-related groups of taxa

Most previous studies used nucleotide or codon models to investigate the non-homogeneity assumption in empirical data. Groussin, et al. (Groussin, et al. 2013) introduced a non-homogeneous model for amino acid datasets that reduces the dimensionality of the observed model using the χ^2 distance between the amino acid frequencies. In this study, we used QMaker to estimate the full empirical amino-acid model with 210 free parameters for each clade. This is a non-homogeneous and non-stationary model since each clade will have its substitution rate matrix and base frequencies. Moreover, since the datasets we used for this study are concatenated data, the ML inference could suffer inconsistency issues due to the discordance between the evolution of the different loci (Bryant and Hahn 2020). To account for some of this heterogeneity in the substitution processes, we used a fully partitioned model with edge-linked substitution rates. However, we acknowledge that our analyses rely on the assumption common to all concatenated analyses, that all loci share a common bifurcating phylogenetic tree. We do not think that this assumption would have any major effects on our conclusions though.

Algorithms for clustering branches according to their substitution processes such as (Jayaswal, Ababneh, et al. 2011; Zhang, et al. 2011; Dutheil, et al. 2012) show that there is always a better-fit model than the most complex model with the maximum number of matrices. Those results are consistent with our results in this study. Although, in contrast to those algorithms which allow each branch in the tree to have its own matrix, our algorithm is less general insofar as it starts frompre-defined clades which we assume that all branches in that clade have the same substitution matrix. This assumption is proved to be reasonable as even with this constraint we still get a simpler model with fewer matrices than the number of pre-defined clades to be the best-fit model for all three datasets. These results hint that when previous knowledge about the evolutionary processes on certain clades is available, using pre-defined clades might be enough to significantly improve model estimations.

Yet, using pre-defined clades which we assume that each of them is homogeneous, one would expect that the fully non-homogeneous models with the maximum number of models will outperform simpler models. But as our results show, this is not the case. One limitation of the current approach that could lead to such results is using SRH models for inferring the different matrices. While the overall model is non-homogeneous and non-stationary (except for the case of the homogeneous model), it is still a time-reversible model. Using non-reversible models to estimate the matrices might tell a different story. Another limitation is that the current algorithm estimates one tree for all clades and not a separate tree for each clade as we would ideally want. This causes the deep branches connecting the clades to be included in the calculation of the joint matrix and therefore in the AIC score.

In a molecular evolution framework, the results of this study confirm that different clades in the tree of life have different evolutionary processes. Moreover, they validate that clades from the same kingdom, phylum or class tend to cluster together. For example, our results from the mammals' class show that in the best-fit model the suborders Yinpterochiroptera and Yangochiroptera have a very similar evolutionary process and therefore can be represented by one homogeneous model for the order Chiroptera. The same goes for the two orders Artiodactyla and Cetacea, they have a very similar evolutionary process and can be represented with a single model which ratifies the theory that there is a strong relationship between these two orders and can be represented under the superorder Cetartiodactyla (Millinkovitch and Thewissen 1997; Montgelard, et al. 1997; Shimamura, et al. 1997; Naylor and Adams 2001; Thewissen, et al. 2001; Theodor and Foss 2005; Agnarsson and May-Collado 2008; Hassanin, et al. 2012).

In conclusion, our study emphasizes the importance of using non-homogeneous models to investigate evolutionary processes. Several non-homogeneous models are available for phylogenetic analysis, yet, they are still rarely used in empirical studies. As our results show, it is worthwhile to invest more in making those methods more accessible and user-friendly and to develop new methods that can be non-stationary, non-homogeneous and non-reversible.

Appendix

The Greedy Algorithm commandlines

1)
```
iqtree2 -s ALIGNMENT_FILE -p PARTITION_FILE --prefix SRH
```

Where ALIGNMENT_FILE is the alignment file with all the taxa used in this study, and PARTITION_FILE is the partition file.

2)
```
iqtree2 -s ALIGNMENT_FILE -p SRH.best_model.nex -te SRH.treefile
--model-joint GTR20+FO --min-freq 0.001 --prefix SRH_Q -nparam 10
-optfromgiven
```

Where SRH.best_model.nex is the best model output from command line 1 and SRH.treefile is the ML tree output from command line 1.

--model-joint GTR20+FO --min-freq 0.001 means use QMaker to estimate the general time-reversible model with 20-state data (GTR20) with ML estimate of the state frequencies (FO) and set the minimum state frequencies to 0.001. -optfromgiven means that we do not fix the model parameters and allow IQ-TREE to optimize them further, and –nparam 10 means that we use 10 iterations to estimate the Q matrix and stationary frequencies vector.

3)
```
iqtree2 -s CLADE_ALIGNMENT_FILE -p PARTITION_FILE --prefix
CLADE_SRH
```

Where CLADE_ALIGNMENT_FILE is the alignment file with all the taxa in a specific clade, PARTITION_FILE is the partition file.

4)

```
iqtree2 -s CLADE_ALIGNMENT_FILE -p CLADE_SRH.best_model.nex -te CLADE_SRH.treefile --model-joint GTR20+FO --min-freq 0.001 --prefix CLADE_Q -nparam 10 -optfromgiven
```

Where CLADE_SRH.best_model.nex is the best model output from command line 3 and CLADE_SRH.treefile is the ML tree output from command line 3.

References

Ababneh F, Jermiin LS, Ma C, Robinson J. 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics 22:1225-1231.

Agnarsson I, May-Collado LJ. 2008. The phylogeny of Cetartiodactyla: the importance of dense taxon sampling, missing data, and the remarkable promise of cytochrome b to provide reliable species-level phylogenies. Mol Phylogenet Evol 48:964-985.

Akaike H. 1974. A new look at the statistical model identification. IEEE transactions on automatic control 19:716-723.

Bazinet AL, Zwickl DJ, Cummings MP. 2014. A gateway for phylogenetic analysis powered by grid computing featuring GARLI 2.0. Syst. Biol. 63:812-818.

Blanquart S, Lartillot N. 2006. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Mol. Biol. Evol. 23:2058-2071.

Blanquart S, Lartillot N. 2008. A site- and time-heterogeneous model of amino acid replacement. Mol. Biol. Evol. 25:842-858.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comp. Biol. 10:e1003537.

Boussau B, Blanquart S, Necsulea A, Lartillot N, Gouy M. 2008. Parallel adaptations to high temperatures in the Archaean eon. Nature 456:942-945.

Boussau B, Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. Syst. Biol. 55:756-768.

Bryant D, Hahn MW. 2020. The Concatenation Question. In: Scornavacca C, Delsuc F, Galtier N, editors. Phylogenetics in the Genomic Era: No commercial publisher | Authors open access book. p. 3.4:1--3.4:23.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

Dutheil J, Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. BMC Evol. Biol. 8:255.

Dutheil JY, Galtier N, Romiguier J, Douzery EJ, Ranwez V, Boussau B. 2012. Efficient selection of branch-specific models of sequence evolution. Mol. Biol. Evol. 29:1861-1874.

Foster PG. 2004. Modeling compositional heterogeneity. Syst. Biol. 53:485-495.

Galtier N, Gouy M. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. 15:871-879.

Galtier N, Gouy M. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proc Natl Acad Sci U S A 92:11317-11321.

Galtier N, Tourasse N, Gouy M. 1999. A nonhyperthermophilic common ancestor to extant life forms. Science 283:220-221.

Groussin M, Boussau B, Gouy M. 2013. A branch-heterogeneous model of protein evolution for efficient inference of ancestral sequences. Syst. Biol. 62:523-538.

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

Hassanin A, Delsuc F, Ropiquet A, Hammer C, Jansen van Vuuren B, Matthee C, Ruiz-Garcia M, Catzeflis F, Areskoug V, Nguyen TT, et al. 2012. Pattern and timing of diversification of Cetartiodactyla (Mammalia, Laurasiatheria), as revealed by a comprehensive analysis of mitochondrial genomes. C. R. Biol. 335:32-50.

Herbeck JT, Degnan PH, Wernegreen JJ. 2005. Nonhomogeneous model of sequence evolution indicates independent origins of primary endosymbionts within the enterobacteriales (gamma-Proteobacteria). Mol. Biol. Evol. 22:520-532.

Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, Huelsenbeck JP, Ronquist F. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst. Biol. 65:726-736.

Jarvis ED, Mirarab S, Aberer A, Houde P, Li C, Ho S, Faircloth BC, Nabholz B, Howard JT, Suh A, et al. 2014. Data from: Phylogenomic analyses data of the avian phylogenomics project. In: GigaScience Database.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT, et al. 2015. Phylogenomic analyses data of the avian phylogenomics project. Gigascience 4:4.

Jayaswal V, Ababneh F, Jermiin LS, Robinson J. 2011. Reducing model complexity of the general Markov model of evolution. Mol. Biol. Evol. 28:3045-3059.

Jayaswal V, Jermiin LS, Poladian L, Robinson J. 2011. Two stationary nonhomogeneous Markov models of nucleotide sequence evolution. Syst. Biol. 60:74-86.

Jayaswal V, Jermiin LS, Robinson J. 2005. Estimation of Phylogeny Using a General Markov Model. Evol Bioinform 1:62-80.

Jayaswal V, Robinson J, Jermiin L. 2007. Estimation of phylogeny and invariant sites under the general Markov model of nucleotide sequence evolution. Syst. Biol. 56:155-162.

Jayaswal V, Wong TK, Robinson J, Poladian L, Jermiin LS. 2014. Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. Syst. Biol. 63:726-742.

Jermiin L, Ho SY, Ababneh F, Robinson J, Larkum AW. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst. Biol. 53:638-643.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14:587-589.

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.

Millinkovitch MC, Thewissen JG. 1997. Evolutionary biology. Even-toed fingerprints on whale ancestry. Nature 388:622-624.

Minh BQ, Dang CC, Vinh LS, Lanfear R. 2021. QMaker: Fast and Accurate Method to Estimate Empirical Models of Protein Evolution. Syst. Biol. 70:1046-1060.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. 37:1530-1534.

Montgelard C, Catzeflis FM, Douzery E. 1997. Phylogenetic relationships of artiodactyls and cetaceans as deduced from the comparison of cytochrome b and 12S rRNA mitochondrial sequences. Mol. Biol. Evol. 14:550-559.

Naser-Khdour S, Lanfear R, Minh BQ. 2021. The Influence of Model Violation on Phylogenetic Inference: A Simulation Study. bioRxiv:2021.2009.2022.461455.

Naser-Khdour S, Minh BQ, Zhang W, Stone EA, Lanfear R. 2019. The Prevalence and Impact of Model Violations in Phylogenetic Analysis. Genome Biol Evol 11:3341–3352.

Naylor GJ, Adams DC. 2001. Are the fossil data really at odds with the molecular data/morphological evidence for Cetartiodactyla phylogeny reexamined. Syst. Biol. 50:444-453.

Nguyen LT, 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.

Ran JH, Shen TT, Wang MM, Wang XQ. 2018a. Data from: Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between Gnetales and angiosperms. In: Dryad Digital Repository.

Ran JH, Shen TT, Wang MM, Wang XQ. 2018b. Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between Gnetales and angiosperms. Proc Biol Sci 285:20181012.

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

Shimamura M, Yasue H, Ohshima K, Abe H, Kato H, Kishiro T, Goto M, Munechika I, Okada N. 1997. Molecular evidence from retroposons that whales form a clade within even-toed ungulates. Nature 388:666-670.

Song H, Sheffield NC, Cameron SL, Miller KB, Whiting MF. 2010. When phylogenetic assumptions are violated: base compositional heterogeneity and among-site rate variation in beetle mitochondrial phylogenomics. Syst. Entomol. 35:429-448.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312-1313.

Swofford DL. 2001. Paup*: Phylogenetic analysis using parsimony (and other methods) 4.0. B5.

Theodor JM, Foss SE. 2005. Deciduous dentitions of Eocene cebochoerid artiodactyls and cetartiodactyl relationships. J. Mamm. Evol. 12:161-181.

Thewissen JG, Williams EM, Roe LJ, Hussain ST. 2001. Skeletons of terrestrial cetaceans and the relationship of whales to artiodactyls. Nature 413:277-281.

Wu S, Edwards S, Liu L. 2019. Data from: Genome-scale DNA sequence data and the evolutionary history of placental mammals. In: Figshare.

Wu S, Edwards S, Liu L. 2018. Genome-scale DNA sequence data and the evolutionary history of placental mammals. Data Brief 18:1972-1975.

Yang Z. 2006. Computational Molecular Evolution. Oxford, UNITED KINGDOM: Oxford University Press USA - OSO.

Yang ZH, Roberts D. 1995. On the Use of Nucleic-Acid Sequences to Infer Early Branchings in the Tree of Life. Mol. Biol. Evol. 12:451-458.

Zhang C, Wang J, Xie W, Zhou G, Long M, Zhang Q. 2011. Dynamic programming procedure for searching optimal models to estimate substitution rates based on the maximum-likelihood method. Proc Natl Acad Sci U S A 108:7860-7865.

Zou L, Susko E, Field C, Roger AJ. 2012. Fitting nonstationary general-time-reversible models to obtain edge-lengths and frequencies for the Barry–Hartigan model. Syst. Biol. 61:927-940.

CHAPTER 4

ASSESSING CONFIDENCE IN ROOT PLACEMENT ON PHYLOGENIES: AN EMPIRICAL STUDY USING NON-REVERSIBLE MODELS FOR MAMMALS

Suha Naser-Khdour*[1], Bui Quang Minh[1,2], and Robert Lanfear[1]

(7)Department of Ecology and Evolution, Research School of Biology, Australian National University, Canberra, Australian Capital Territory, Australia

(8) Research School of Computer Science, Australian National University, Canberra, Australian Capital Territory, Australia

**Author for Correspondence: E-mail:* suha.naser@anu.edu.au

Contributions:

SNK wrote the python script, performed the analysis, analysed and interpreted the results, drafted the manuscript, and submitted the article for publication. MB contributed to the research design, conceptual development, editorial comments and implemented the metrics in IQ-TREE. RL contributed to the research design, conceptual development and editorial comments.

Abstract

Using time-reversible Markov models is a very common practice in phylogenetic analysis, because although we expect many of their assumptions to be violated by empirical data, they provide high computational efficiency. However, these models lack the ability to infer the root placement of the estimated phylogeny. In order to compensate for the inability of these models to root the tree, many researchers use external information such as using outgroup taxa or additional assumptions such as molecular-clocks. In this study, we investigate the utility of non-reversible models to root empirical phylogenies and introduce a new bootstrap measure, the *rootstrap*, which provides information on the statistical support for any given root position.

Availability and implementation: rootstrap support is implemented in IQ-TREE 2 and a tutorial is available at the iqtree webpage http://www.iqtree.org/doc/Rootstrap. In addition, a python script is available at https://github.com/suhanaser/Rootstrap.

[phylogenetic inference, root estimation, bootstrap, non-reversible models]

Main Text

The most widely used method for rooting trees in phylogenetics is the outgroup method. Although the use of an outgroup to root an unrooted phylogeny usually outperforms other rooting methods (Huelsenbeck, et al. 2002), the main challenge with this method is to find an appropriate outgroup (Watrous and Wheeler 1981; Maddison, et al. 1984; Smith 1994; Swofford, et al. 1996; Lyons-Weiler, et al. 1998; Milinkovitch and Lyons-Weiler 1998). Outgroups that are too distantly-related to the ingroup may have substantially different molecular evolution than the ingroup, which can compromise accuracy. And outgroups that are too closely related to the ingroup may not be valid outgroups at all.

It is possible to infer the root of a tree without an outgroup using molecular clocks (Huelsenbeck, et al. 2002; Drummond, et al. 2006). A strict molecular-clock assumes that the substitution rate is constant along all lineages, a problematic assumption especially when the ingroup taxa are distantly related such that their rates of molecular evolution may vary. Relaxed molecular-clocks are more robust to deviations from the clock-like behaviour (Drummond, et al. 2006), although previous studies have shown that they can perform poorly in estimating the root of a phylogeny when those deviations are considerable (Tria, et al. 2017).

Other rooting methods rely on the distribution of branch lengths, including Midpoint Rooting (MPR) (Farris 1972), Minimal Ancestor Deviation (MAD) (Tria, et al. 2017), and Minimum Variance Rooting (MVR) (Mai, et al. 2017). Such methods also assume a clock-like behaviour; however, they are less dependent on this assumption as the unrooted tree is estimated without it. Similar to inferring a root directly from molecular-clock methods, the accuracy of those rooting methods decreases with higher deviations from the molecular-clock assumption (Mai, et al. 2017).

Other less common rooting methods that can be used in the absence of outgroup are: rooting by gene duplication (Dayhoff and Schwartz 1980; Gogarten, et al. 1989; Iwabe, et al. 1989), indel-based rooting (Rivera and Lake 1992; Baldauf and Palmer 1993; Lake, et al. 2007), rooting the species tree from the distribution of unrooted gene trees (Allman, et al. 2011; Yu, et al. 2011), and probabilistic co-estimation of gene trees and species tree (Boussau, et al. 2013).

All the methods mentioned above, apart from the molecular-clock, infer the root position independently of the ML tree inference. The only existing approach to include root placement in the ML inference is the application of non-reversible models. Using non-reversible substitution models relaxes the fundamental assumption of time-reversibility that exists in the most widely used models in phylogenetic inference (Jukes and Cantor 1969; Kimura 1980; Hasegawa, et al. 1985; Tavaré 1986; Dayhoff 1987; Jones, et al. 1992; Tamura and Nei 1993; Whelan and Goldman 2001; Le and Gascuel 2008). This in itself is a potentially useful improvement in the fit between models of sequence evolution and empirical data. In addition, since non-reversible models naturally incorporate a notion of time, the position of the root on the tree is a parameter that is estimated as part of the ML tree inference. Since the incorporation of non-reversible models in efficient ML tree inference software is relatively new (Minh, et al. 2020), we still understand relatively little about the ability of non-reversible models to infer the root of a phylogenetic tree, although a recent simulation study has shown some encouraging results (Bettisworth and Stamatakis 2020).

Regardless of the rooting method and the underlying assumptions, it is crucial that we are able to estimate the statistical confidence we have in any particular placement of the root on a phylogeny. A number of previous studies have sensibly used ratio likelihood tests such as the Shimodaira-Hasegawa (SH) test (Shimodaira and Hasegawa 1999) and the Approximately

Unbiased (AU) test (Shimodaira 2002) to compare a small set of potential root placements, rejecting some alternative root placements in favour of the ML root placement e.g.(Nardi, et al. 2003; Steenkamp, et al. 2006; Jansen, et al. 2007; Moore, et al. 2007; Williams, et al. 2010; Kocot, et al. 2011; Zhou, et al. 2011; Whelan, et al. 2015; Zhang, et al. 2018), these tests are still somewhat limited in that they do not provide the level of support the data have for a certain root position.

There is strong empirical evidence that molecular evolutionary processes are rarely reversible (Squartini and Arndt 2008; Naser-Khdour, et al. 2019), but few studies have explored the accuracy of non-reversible substitution models to root phylogenetic trees (Huelsenbeck, et al. 2002; Yap and Speed 2005; Williams, et al. 2015; Cherlin, et al. 2018; Bettisworth and Stamatakis 2020). Most studies that have looked at this question in the past have focused on either simulated datasets (Huelsenbeck, et al. 2002; Jayaswal, et al. 2011; Cherlin, et al. 2018) or relatively small empirical datasets (Yang and Roberts 1995; Yap and Speed 2005; Jayaswal, et al. 2011; Heaps, et al. 2014; Williams, et al. 2015; Cherlin, et al. 2018). In both cases, the addressed substitution models were nucleotide models, and to our knowledge, no study has yet investigated the potential of amino acid substitution models in inferring the root placement of phylogenetic trees.

In this paper, we focus on evaluating the utility of non-reversible amino acid and nucleotide substitution models to root the trees, and we introduce a new metric, the *rootstrap support value*, which estimates the extent to which the data support every possible branch as the placement of a root in a phylogenetic tree. Unlike previous studies that used Bayesian methods with non-reversible substitution models to infer rooted ML trees (Heaps, et al. 2014; Cherlin, et al. 2018), we will conduct our study in a Maximum likelihood framework using IQ-TREE (Minh, et al. 2020). A clear advantage of Maximum likelihood over the Bayesian analysis is that there is no need for a prior on the parameter distributions, which sometimes can affect tree inference (Huelsenbeck, et al. 2002; Cherlin, et al. 2018). Even though estimating the non-reversible model's parameters by maximizing the likelihood function seems more computationally intensive than calculating posterior probabilities (Huelsenbeck, et al. 2002), the IQ-TREE algorithm is sufficiently fast to allow us to estimate root placements, with *rootstrap support* for very large datasets.

A recent study investigated the ability of non-reversible nucleotide models to infer the root placement of phylogenetic trees (Bettisworth and Stamatakis 2020). This study showed that IQ-TREE performs competitively with a new rooting tool, RootDigger. In most simulated datasets, IQ-TREE slightly outperformed RootDigger in terms of root placements, but no comparisons were made between RootDigger and IQ-TREE on empirical datasets. Although RootDigger is significantly faster than IQ-TREE (Bettisworth and Stamatakis 2020), the former is limited to nucleotide substitution models. Since we are interested in both nucleotide and amino acid non-reversible models, we used IQ-TREE for tree and root inference in this study.

Material and Methods

The "Rootstrap" Support, and measurements of error in root placement

To compute rootstrap supports, we conduct a bootstrap analysis, i.e., resampling alignment sites with replacement, to obtain a number of bootstrap trees. We define the *rootstrap* support for each branch in the ML tree, as the proportion of bootstrap trees that have the root on that branch. Since the root can be on any branch in a rooted tree, the rootstrap support values are computed for all the branches including external branches. The sum of the rootstrap support values along the tree is always smaller than or equal to one. A sum that is smaller than one can occur when one or more bootstrap replicates are rooted on a branch that does not occur in the ML tree (Fig. 1).

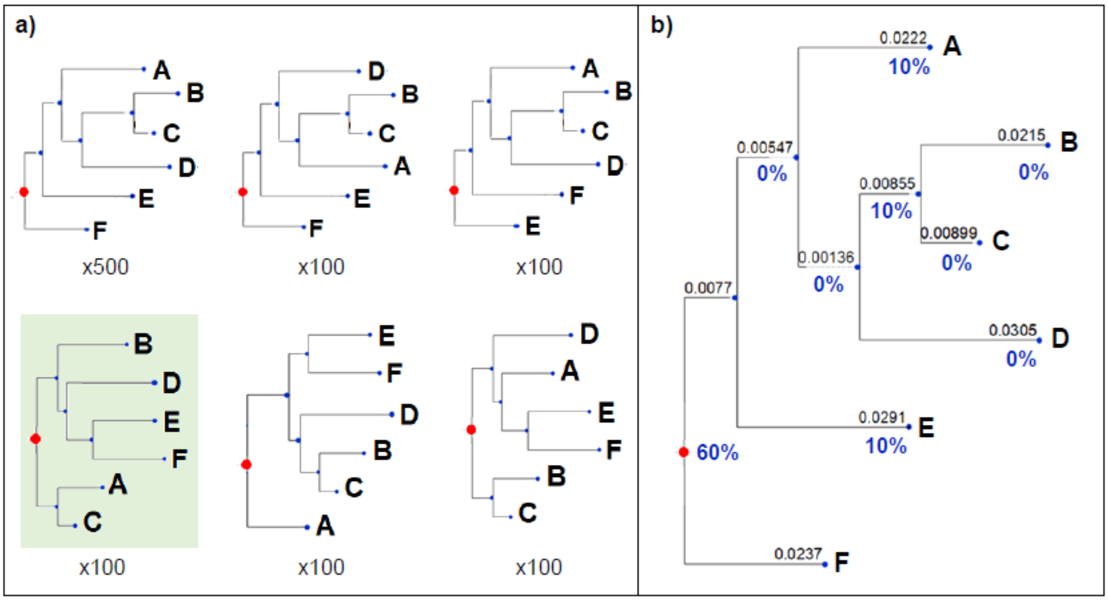


FIGURE 1. Illustration of the rootstrap concept. (a) The bootstrap replicates trees. (b) The ML tree with the rootstrap support values for each branch. Note that the sum of the rootstrap support values is less than 100% due to 100 bootstrap replicates trees (green) that have their root at a branch that does not exist in the ML tree.

By definition, the rootstrap support values for internal branches are bounded by the bootstrap support values at those branches. On the other hand, the rootstrap support values for tips (leaf branches) are bounded by 100%, as tips always appear in all the bootstrap trees.

If the true position of the root is known (e.g. in simulation studies) or assumed (e.g. in the empirical cases we present below), we can calculate additional measurements of the error of the root placement. We introduce two such measurements here: *root branchlength error distance* (rBED) and *root split error distance* (rSED). Since the non-reversible model infers the exact position of the root on a branch, we define the *root branchlength error distance* (rBED) as the range between the minimum and maximum distance between the inferred root position and the "true root" branch. If the true root is on the same branch as the ML tree root, then rBED will be between 0 and the distance between the ML tree root and the farthest point on that branch (Fig. 2). Since rBED is based on branch lengths only, it ignores the absolute number of splits between the ML tree root and the true root; and therefore, the rBED for the true root being on the same ML root branch can be bigger than the rBED for the true root being on a different branch (e.g. Fig. 2). In order to account for the number of splits (nodes) between the ML tree root and the true root, we define *root split error distance* (rSED) as the number of splits between the ML root branch and the branch that is believed to contain the true root (Fig. 2).

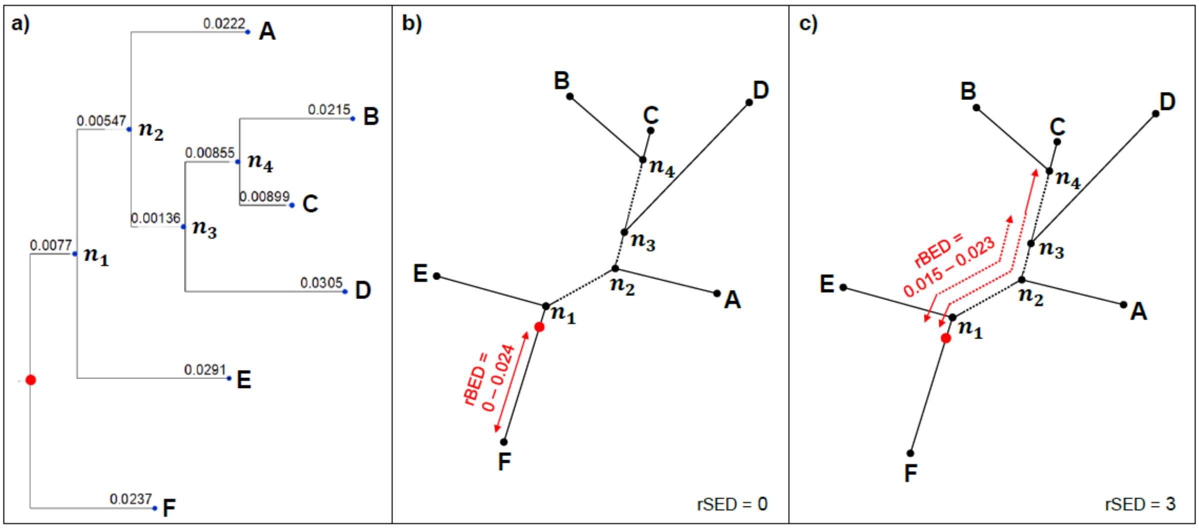


FIGURE 2. An example to illustrate the root error distance. (a) the ML rooted tree, **(b)** the root branch-length error distance (rBED) if the true root is believed to be on the same ML root branch (rSED = 0), **(c)** the rBED if the true root is believed to be on the branch between D and the clade of C + B (rSED = 3).

The rootstrap, rBED, and rSED assess different aspects of the root placement. While the rootstrap offers an indication of the support that the data have for a certain branch to be the root branch, rBED and rSED provide an estimation of the accuracy of the method in estimating the exact root position if the root position is known or assumed in advance. In other words, the rootstrap value is a measure for the robustness of the root placement given the model and the data and can be used on any dataset regardless of whether the true root position is known, while rBED and rSED are measures of the accuracy of the non-reversible model to find the root placement given the data, and require the root position to be known or assumed in advance.

Empirical Datasets

Because non-reversible amino acid models require the estimation of a large number of parameters, and because we suspected that the information in any such analysis on the placement of the root branch of a tree might be rather limited, we searched for empirical datasets that met a number of stringent criteria:

1) Existence of both DNA and amino acid multiple sequences alignments (MSA) for the same loci.
2) Genome-scale MSAs to ensure that the MSAs have as much information as possible with which to estimate the non-reversible models' free parameters and the root position. Since we do not know the number of sites required to correctly infer the rooted ML tree, we define 100,000 sites as the minimum number of required sites. This also allows us to subsample the dataset to explore the ability of smaller datasets to infer root positions.
3) Highly-curated alignments: since the quality of the inferred phylogeny is highly dependent on the quality of the MSA (Philippe, et al. 2011), we focussed on datasets that were highly-curated for misalignment, contamination, and paralogy.
4) Existence of several clades for which there is a very strong consensus regarding their root placement. Since we are interested in evaluating the performance of non-reversible models to infer root placements in an empirical rather than a simulation context, we need to identify monophyletic sub-clades for which we can be almost certain about their root position. This enables us to divide the dataset into non-overlapping sub-clades for

which we are willing to assume we know the root positions. Furthermore, we define the minimum number of taxa in each sub-dataset as five.

We initially identified a number of genome-scale datasets that contained large numbers of nucleotide and amino acid MSAs. In many cases, it was difficult to determine whether these alignments had been rigorously curated, and even more challenging to find datasets for which the root position of a number of subclades could be assumed with confidence. The only dataset that met all of our criteria was a dataset of placental mammals with 78 ingroup taxa and 3,050,199 amino acids (Wu, et al. 2019). This dataset was originally published as an MSA (Liu, et al. 2017) based on very high-quality sequences from Ensembl, NCBI, and GenBank databases. After receiving detailed critiques for potential alignment errors (Gatesy and Springer 2017), the dataset was further processed to remove potential sources of bias and error, and an updated version of the dataset was recently published (Wu, et al. 2018). The fact that this alignment comes from one of the most well-studied clades on the planet, has been independently curated and critiqued by multiple groups of researchers and includes truly genome-scale data, makes it ideally suited for our study. The curated alignments can be found on figshare (https://figshare.com/s/622e9e0a156e5233944b) under the name "Wu_2018_aa" and "Wu_2018_dna" for the amino-acid and nucleotide alignments respectively.

Selecting Clades with a Well-Defined Root

Since our main objective in this study is to evaluate the effectiveness of non-reversible models and the rootstrap value in estimating and measuring the support for a given root placement on empirical datasets, we must identify a collection of sub-clades of the larger mammal dataset for which it is reasonable to assume a root position. We acknowledge, of course, that outside a simulation framework it is not possible to be certain of the root position of a clade. Nevertheless, it is possible to identify clades for which the position of the root is well supported and non-controversial, thus minimising the chances that the assumption of a particular root position is incorrect. To achieve this, we analysed the root position of each order and superorder in the dataset, and defined "*well-defined clades*" that fulfilled **all** of the following criteria:

(1) It contains at least five taxa. This ensures that the probability of obtaining a random ML rooted tree to be at most 0.95%. For clades with four taxa, there are 15 different rooted topologies, and therefore a 6.7% probability to get any particular root position by chance. On the other hand, for clades with at least five taxa, there are at least 105 different rooted topologies and maximum probability of 0.95% to randomly get a particular root position by chance.

(2) The bootstrap support for the branch leading to that clade in the phylogenetic tree calculated from the whole dataset is 100%: since the bootstrap value indicates the support the data have for a certain branch, we also require 100% support for the first direct descendants in the clade (Appendix Fig. A.1). This requirement ensures that there is strong support in the dataset for the root position of the clade when the entire dataset is rooted with an outgroup.

(3) The site concordance factor (sCF) for the first direct descendants in the clade is significantly greater than 33%. The site Concordance Factor (sCF) is calculated by comparing the support of each site in the alignment for the different arrangements of quartet around a certain branch. In other words, an sCF of 33% means equal support for any of the possible arrangements. Therefore, we require that the sCF of the deepest two levels of branches leading to that clade is significantly greater than 33%. Moreover, we require that the gene Concordance Factor (gCF) for the first direct descendants in the clade to be significantly greater than 33% of the sum of the gene concordance factor and the two Discordance Factors (gDF1 and gDF2). The gCF of a branch is calculated as the proportion of gene trees containing that branch, and gDFs are calculated as the proportion of gene

trees containing one of the two other resolutions of that branch. Since for each branch in a bifurcating tree there are three possible arrangements of clades around that branch, we ignore all gene trees that do not contain one of these arrangements (e.g. gene trees that contribute to neither the gCF nor the gDFs). Although there is no threshold regarding the required proportion of genes concordant with a certain branch, for convenience, we define branches with gCF significantly greater than 33% of the sum gCF+gDF1+gCF2 as branches that are concordant with enough genes in the alignment (Minh, et al. 2020). To test whether the sCF and the gCF are significantly greater than 33%, we use a simple binomial test with a success probability of 0.33. The gCF,gDF1,gCF2 and sCF values are based on the tree estimated from the amino acid dataset.

(4) At least 95% of the studies that have been published in the last decade support this clade: we searched google scholar for all published papers since 2009 that determine the root of the addressed clade. We then checked if at least 95% of those papers agree that the root position of the clade matches that in the ML tree we estimate from the whole dataset (see supplementary material).

Estimating unrooted Phylogenies

For the whole nucleotide and amino-acid datasets with ingroup and outgroup taxa, we inferred the unrooted phylogeny using IQ-TREE2 (Minh, et al. 2020) with the best-fit fully partitioned model (Chernomor, et al. 2016) and edge-linked substitution rates (Duchene, et al. 2020). We then determined the best-fit reversible model for each partition using ModelFinder (Kalyaanamoorthy, et al. 2017). See the algorithm for finding well-defined clades in Appendix Algorithm A.1.

Estimating Rooted phylogenies

For each well-defined clade, we first removed all other taxa from the tree and then sought to infer the root of the well-defined clade using non-reversible models without outgroups. Using the best partitioning scheme from the reversible analysis, we inferred the rooted tree for each well-defined clade with the non-reversible models for amino acid (NR-AA) and nucleotide (NR-DNA) sequences (Minh, et al. 2020). This approach fits a 12-parameter non-reversible model for DNA sequences, and a 380-parameter non-reversible model for amino acids. Details of the command lines used are provided in the supplementary material section "Algorithm A.2". Each analysis returns a rooted tree. We performed 1000 non-parametric bootstraps of every analysis to measure the rootstrap support.

To assess the performance of the rootstrap and the ability of non-reversible models to estimate the root of the trees on smaller datasets, we also repeated every analysis on subsamples of the complete dataset. For each well-defined clade, we performed analysis on the complete dataset (100%) as well as datasets with 10%, 1% and 0.1% of randomly-selected loci from the original alignment.

The confidence set of root branches using the Approximately Unbiased test

In addition to the rootstrap support, we calculate the confidence set of all the branches that may contain the root of the ML tree using the Approximately Unbiased (AU) test (Shimodaira 2002). To do this, we re-root the ML tree with all possible placements of the root (one placement for each branch) and calculate the likelihood of each tree. Using the AU test, we then ask which root placements can be rejected in favour of the ML root, using an alpha value of 5%. We define the *root branches confidence set* as the set of root branches that are not rejected in favour of the ML root placement. An important difference between the AU test and the rootstrap support is that the AU test is conditioned on a single ML tree topology, but the rootstrap support is not. Because of this, they provide quite different information about the position of the root. The AU test assumes that the ML tree topology is true and then seeks to determine the confidence set of root placements conditioned on that topology. The confidence set for the AU test will always, therefore, contain at least the ML root branch. The rootstrap does not assume any particular topology and instead asks how many times a particular root position appears across a set of bootstrap replicates. Because of this, it is possible for every branch in the ML topology to receive 0% rootstrap support. This can occur if none of the branches in the ML topology appears as the root branch in any of the bootstrap topologies.

Reducing systematic bias by removing third codon positions and loci that fail the MaxSym test

As it is common in many phylogenetic analyses to remove third codon positions from the alignment (Swofford, et al. 1996), we wanted to assess the effect of removing third codon positions on the root inference and the rootstrap values in nucleotide datasets. For that purpose, we remove all the third codon positions from the nucleotide alignments and re-ran the analysis using the NR-DNA model.

Moreover, although the NR-AA and NR-DNA models relax the reversibility assumption, they still assume stationarity and homogeneity. To reduce the systematic bias produced by violating these assumptions, we used the MaxSym test (Naser-Khdour, et al. 2019) to remove loci that violate those assumptions in the nucleotide and amino acid datasets, and then re-ran all analyses as above.

Applying the methods to two clades whose root position is uncertain

In addition to the well-defined clades, we used the methods we propose here to infer the root of two clades of mammals whose root position is controversial; Chiroptera and the Cetartiodactyla.

There is a controversy around the root of the Chiroptera (bats) in literature. The two most popular hypotheses are: 1) the Microchiroptera-Megachiroptera hypothesis; where the root is placed between the Megachiroptera, which contains the family Pteropodidae, and the Microchiroptera, which contains all the remaining Chiroptera families. This hypothesis is well supported in the literature (Agnarsson, et al. 2011; Meredith, et al. 2011). However, more recent studies seem to provide less support for this hypothesis; 2) the Yinpterochiroptera-Yangochiroptera hypothesis, in which the Yangochiroptera clade includes most of Microchiroptera and the Yinpterochiroptera clade includes the rest of Microchiroptera and all of Megachiroptera. There is growing support for this hypothesis in the literature (Meganathan, et al. 2012; Tsagkogeorga, et al. 2013; Ren, et al. 2018; Reyes-Amaya and Flores 2019).

Similar to Chiroptera, the root of Cetartiodactyla remains contentious in the literature. The three main hypotheses regarding the root of Cetartiodactyla are: 1) Tylopoda as the sister group for all other cetartiodactylans; 2) Suina as the sister group for all other cetartiodactylans; 3) the monophyletic clade containing Tylopoda and Suina as the sister group for all other cetartiodactylans.

To ascertain whether certain sites or loci had very strong effects on the placement of the root we follow the approach of Shen et. al. (Shen, et al. 2017) and calculate the difference in site-wise log-likelihood scores (ΔSLS) and gene-wise log-likelihood scores (ΔGLS) between the supported root positions for each clade. Moreover, we analysed subsamples of each dataset to test the limits of using non-reversible models to root trees with smaller datasets.

Results

Inference of the mammal tree and selection of well-defined clades

The trees inferred from the whole datasets with the nucleotide-reversible model and the amino-acid-reversible model (Appendix Fig. A.2, Appendix Fig. A.3, Appendix Table A.2) are consistent with the published tree (Liu, et al. 2017). Five clades met all the criteria of well-defined clades, namely, Afrotheria, Bovidae, Carnivora, Myomorpha, and Primates in both amino acid and nucleotide datasets (see Appendix Table A.1 and Appendix Table A.2). Trees in Newick format can be found on github:

https://github.com/suhanaser/Rootstrap/tree/master/trees

High accuracy of the AA non-reversible model in inferring the root

Using NR-AA, we inferred the correct root with very high rootstrap support for all five well-defined clades when all loci were used (Appendix Table A.3). Moreover, for all the five clades, the true root was the only root placement in the confidence set of the AU test. The average running time of the NR-AA model (model estimation + tree search + bootstrap + root inference) is 929 hrs on one core 2.6GHz CPU. However, using the optimal number of cores for each dataset reduced the average running time to 43.5 hrs per dataset.

Our results show that using only 10% of the sites in the amino acid alignments (around 300,000 alignment columns) still gave very high rootstrap support values (> 98%) for four of the five well-defined clades (Fig. 3) with no correlation between rSED and rBED and the size of the dataset (Table A.3). Moreover, in three of five well-defined clades, 1% of the sites (around 30,000 alignment columns) was enough to give a very high rootstrap support value for the assumed correct root placement. Using only 0.1% of the sites (around 3000 alignment columns) decreased the rootstrap support value noticeably in all datasets (Appendix Table A.3). These values are shown for each dataset in Figure 3, where the X-axis is plotted in terms of parsimony-informative sites to allow for a more direct comparison between datasets, and to assist those applying these methods in deciding whether to use them on their own data. Although the rootstrap support for the true root improves as the number of parsimony-informative sites increase, in some datasets (e.g. Afrotheria nucleotide dataset) this is not the case (Fig. 3).

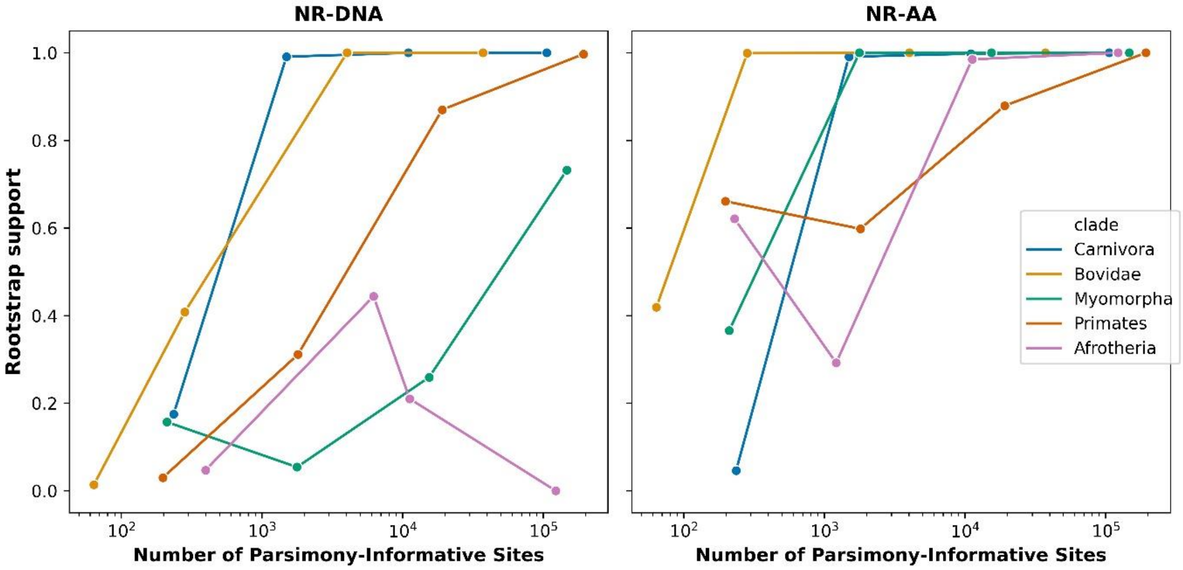


FIGURE 3. The rootstrap support value for each clade as a function of the number of parsimony-informative sites.

The non-reversible amino acid models were strongly preferred to the reversible models on the complete datasets (BIC values were 93943 to 235958 units better for the non-reversible models), and for the datasets with 10% of loci subsampled (BIC values were 3577 to 15082 units better for the non-reversible models), but the opposite was true for the datasets 1% and 0.1% of the loci subsampled (e.g. BIC values were between 2102 and 2712 units worse for the non-reversible models for the 0.1% subsampled datasets; see Table A.7 for full results).

Poor performance of the DNA non-reversible model in inferring the root

We correctly inferred the root for four out of the five nucleotide datasets with the NR-DNA model, when all loci were used. However, the rootstrap support was generally lower than in the amino-acid datasets (Fig. 3, Appendix Tables A.3 and A.4). Similar to amino-acid datasets, there is no correlation between rSED and rBED and the size of the dataset (Table A.4). The average running time of the NR-DNA model (model estimation + tree search + bootstrap + root inference) is 35.7 hrs on one core 2.6GHz CPU and 4 hours when the optimal number of

In contrast to the NR-AA model, there is no conclusive preference for the NR-DNA model over the reversible DNA model for the datasets we analysed (Table A.8). In fact, the BIC values of the NR-DNA models are always worse than reversible models regardless of the size of the nucleotide dataset except for three clades when all loci were included (Table A.8). In two of the datasets (Myomorpha and Primates) where the NR-DNA model was better than the reversible model, the root placement was inferred correctly with high rootstrap support (>95%). In fact, the Afrotheria nucleotide dataset is the only dataset in which the non-reversible model was better than the reversible model but the root placement was inferred incorrectly.

cores for each dataset were used.

TABLE 1. Rootstrap support and rSED values in whole nucleotide datasets and nucleotide datasets without third codon positions.

Clades	All loci		Without 3rd	
	rootstrap	rSED	rootstrap	rSED
Afrotheria	0.0%	2	0.0%	2
Primates	99.7%	0	90.1%	0
Myomorpha	73.2%	0	15.8%	1
Carnivora	100.0%	0	100.0%	0
Bovidae	100.0%	0	82.5%	0

Removing loci that violate the stationarity and homogeneity assumptions improves

Our results show that removing the third codon positions does not improve the rootstrap support value. In contrast, in some datasets removing third codon positions decreased the rootstrap support value and increased the rSED (Table 1).

the rootstrap support

As expected, our results show that removing loci that fail the MaxSym test improves the rootstrap support values when the rootstrap support value was less than 100% and/or the root placement was inferred incorrectly, as the case in some nucleotide datasets (Table 2).

TABLE 2. Rootstrap support values in whole datasets and datasets with loci that passed the MaxSym test only.

	Amino Acid		Nucleotide	
Clade	all loci	Passed MaxSym	all loci	Passed MaxSym
Afrotheria	100.0%	100.0%	0.0%	8.4%
Primates	100.0%	100.0%	99.7%	99.9%
Myomorpha	100.0%	100.0%	73.2%	88.3%
Carnivora	100.0%	100.0%	100.0%	100.0%
Bovidae	100.0%	100.0%	100.0%	100.0%

Microchiroptera-Megachiroptera or Yinpterochiroptera-Yangochiroptera?

Using the whole amino acid dataset, our results show 65.5% rootstrap support for the Yinpterochiroptera-Yangochiroptera hypothesis and 23.2% for the Microchiroptera - Megachiroptera hypothesis. The remaining11.3% of the rootstrap support goes to supporting the branch leading to Rhinolophoidea as the root branch of the bats (Fig. 4). Removing amino acid loci that fail the MaxSym test (110 loci) gives similar results, with 65.9% rootstrap support for the Yinptero-Yango hypothesis and 25.6% rootstrap support for the Micro-Mega hypothesis. In both cases, the AU test could not reject any of the three root positions that received non-zero rootstrap support (Appendix Table A.5).

Using the NR-DNA model gives 100% rootstrap support for the Yinptero-Yango hypothesis, and we can confidently reject the Micro-Mega hypothesis in favour of the Yinptero-Yango hypothesis using the AU test (Appendix Fig. A.4). Yet, removing nucleotide loci that fail the MaxSym test (~25% of the loci) decreases the support for the Yinptero-Yango hypothesis to 90.1%, although we can still confidently reject the Micro-Mega hypothesis using the AU test (Appendix Table A.5).

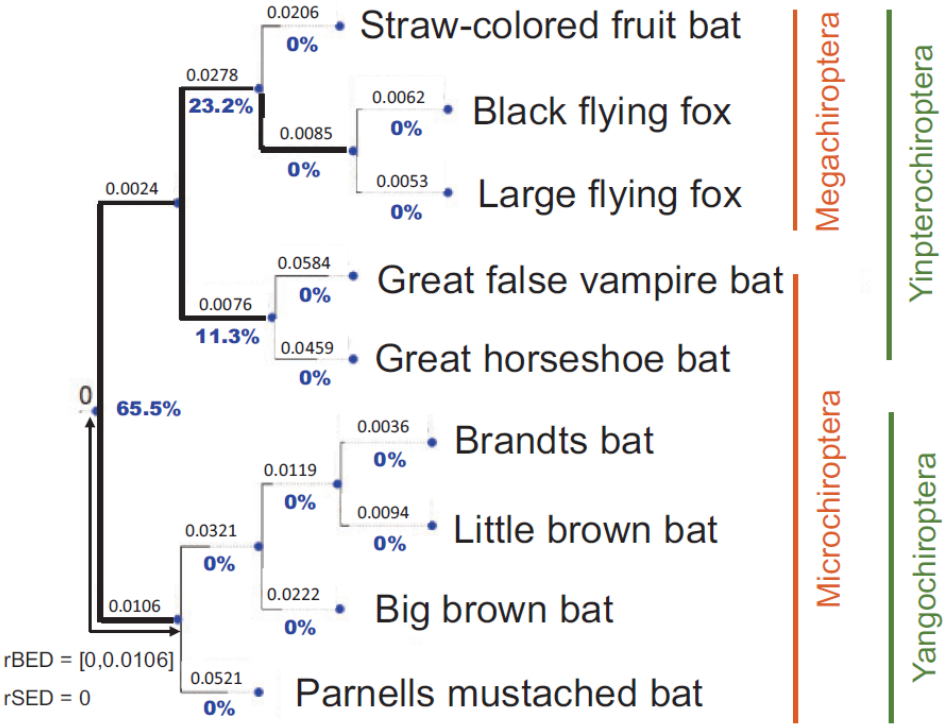


FIGURE 4. The ML rooted tree as inferred from the whole Chiroptera amino acid dataset. Bold branches are branches in the AU confidence set. Blue values under each branch are the rootstrap support values.

Interestingly, when we randomly subsample 10%, 1%, and 0.1% of the loci in the nucleotide dataset, we consistently get the Yinptero-Yango hypothesis as the ML tree and the solely rooted topology in the AU confidence set (Appendix Table A.5). Moreover, the rootstrap support value for the Yinptero-Yango hypothesis increases and the rootstrap support value for the Micro-Mega hypothesis decreases as more parsimony-informative sites are added to the alignment, for both nucleotide and amino acid datasets (Fig. 5, Appendix Table A.5). These results are consistent with previous studies that used smaller datasets (Appendix Figure A.10).

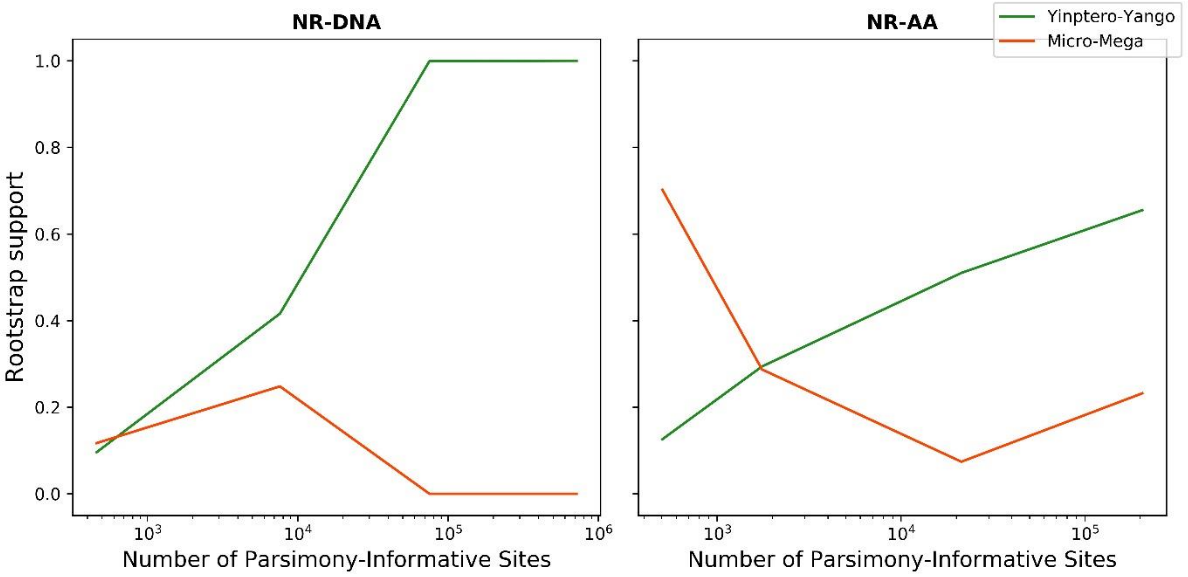


FIGURE 5. Rootstrap support value as a function of the number of parsimony-informative characters in the Chiroptera nucleotide and amino acid datasets using the Non-Reversible DNA model (NR-DNA) and the Non-Reversible Amino Acid model (NR-AA).

The ΔGLS and ΔSLS values (Shen, et al. 2017) reveal that approximately half of the nucleotide and amino acid loci prefer the Yinptero-Yango hypothesis while the other half prefers Micro-Mega hypothesis. Furthermore, slightly less than half of the nucleotide sites prefer the Yinptero-Yango hypothesis. However, more than two-thirds of the amino acid sites prefer the Yinptero-Yango hypothesis (Appendix Fig. A.5). The distributions of ΔGLS and ΔSLS (Appendix Fig. A.6) show that a small proportion of the amino acid loci (~1%) have very strong support for the Micro-Mega hypothesis, and removing those loci from the alignment increased the rootstrap support for the Yinptero-Yango hypothesis to 76.6%. Nonetheless, both root placements are still in the confidence set of the AU test (Appendix Table A.5) with the amino acid dataset. On the other hand, removing nucleotide loci with the highest absolute ΔGLS value still gives the Yinptero-Yango hypothesis as the ML tree and the sole topology in the AU confidence set. Although the nucleotide data show a clear preference to the Yinptero-Yango hypothesis, in terms of BIC scores, the NR-DNA model performs worse than reversible models in all datasets except for the dataset where we removed loci that failed the MaxSym test (Table A.5). On the other hand, the NR-AA performs better than reversible models in big datasets (Table A.5). Yet, the amino acid data do not allow us to distinguish between the two leading hypotheses for the placement of the root of the Chiroptera based on rooting with non-reversible models (Table A.5).

The ambiguous root of Cetartiodactyla

The ML tree inferred with the whole amino acid dataset places the clade containing Tylopoda (represented by its only extant family; Camelidae) and Suina as the sister group to all other cetartiodactylans with 71.8% rootstrap support (Fig. 6). Yet, The AU test did not reject Tylopoda alone as the sister group to all other cetartiodactylans. On the other hand, the ML tree inferred with the whole nucleotide dataset places Tylopoda as the only sister group to all other cetartiodactylans with 71.0% rootstrap support, and we can confidently reject the Tylopoda + Suina hypothesis using the AU test (Appendix Fig. A.7).

Removing the amino acid loci that failed the MaxSym test (~1%) still places Tylopoda + Suina as the sister group to all other cetartiodactylans, yet, it decreases the rootstrap support for the Tylopoda + Suina hypothesis to 63.3% and increases the rootstrap support for the Tylopoda hypothesis to 28.5%. However, we still cannot reject either of the hypotheses using the AU test (Appendix Table A.6).

Removing the nucleotide loci that failed the MaxSym test (~1%) still places Tylopoda as the only sister group to all other cetartiodactylans and the only rooted topology in the AU confidence set. However, it decreases the rootstrap support for the Tylopoda hypothesis to

68.7% and increases the rootstrap support for the Tylopoda + Suina hypothesis to 20.1% (Appendix Table A.6).

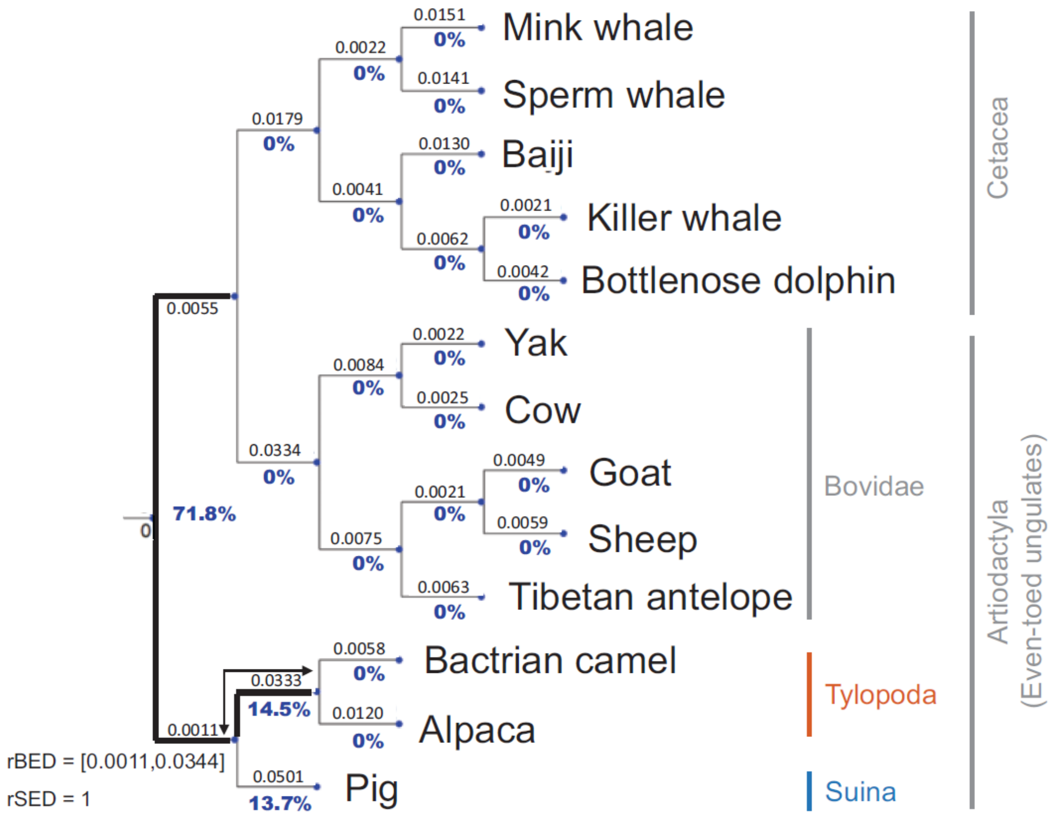


FIGURE 6. The ML rooted tree of as inferred from the whole Cetartiodactyla amino acid dataset. Bold branches are branches in the AU confidence set. Blue values under each branch are the rootstrap support values.

The results from the subsample datasets are mixed (Fig. 7). Analyses on smaller datasets show no clear pattern in the placement of the root (Appendix Table A.6), leading us to conclude only that the analyses of the whole dataset is likely to provide the most accurate result, but that it is plausible that adding more data may lead to different conclusions in the future.

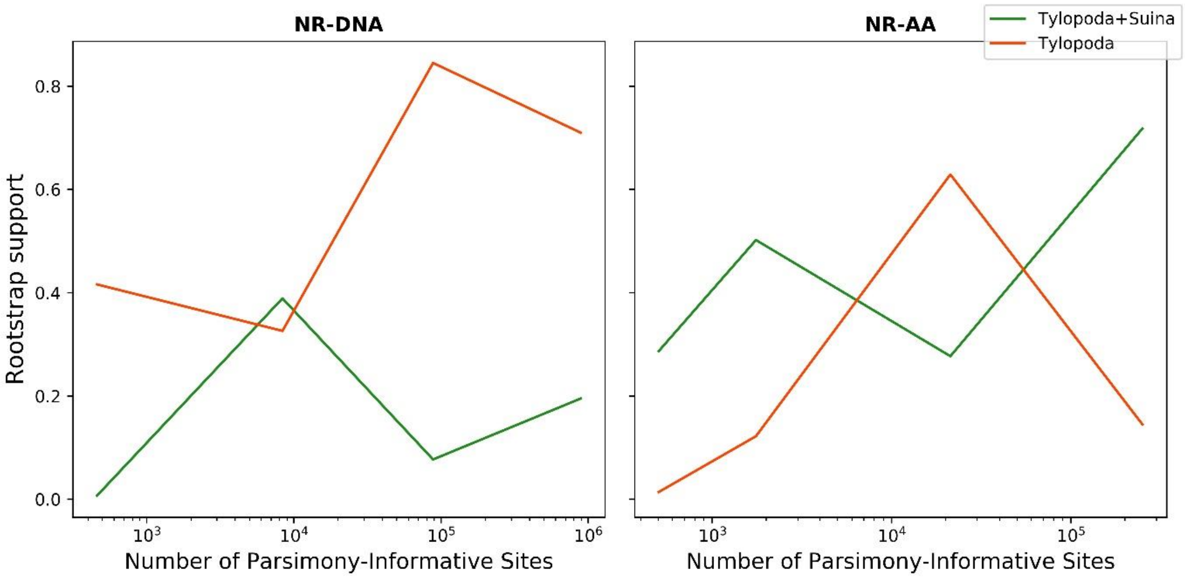


FIGURE 7. rootstrap support value as a function of the number of parsimony-informative characters in the Cetartiodactyla nucleotide and amino acid datasets using the Non-Reversible DNA model (NR-DNA) and the Non-Reversible Amino Acid model (NR-AA).

ΔGLS analyses reveal that approximately, half of the amino acid and nucleotide loci favour the Tylopoda+Suina hypothesis, while the other half of loci favour the Tylopoda hypothesis (Appendix Figs. A.8-9). On the other hand, two-thirds of the amino acid sites and more than 80% of the nucleotide sites favour the Tylopoda+Suina hypothesis. Removing 1% of the amino acid loci with the highest absolute ΔGLS values still places Tylopoda + Suina as the sister group to all other cetartiodactylans. However, the rootstrap support of the Tylopoda + Suina decreased to 63.2% and the rootstrap support for the Tylopoda hypothesis remains approximately the same (~14.5%), while the rootstrap support for the Suina hypothesis increases from 13.7% to 22.4%. Yet, both the Tylopoda + Suina hypothesis and the Tylopoda hypothesis are in the confidence set of the AU test, while the Suina hypothesis is rejected by the AU test (Appendix Table A.6).

Removing 1% of the nucleotide loci with the highest absolute ΔGLS values gives the Tylopoda+Suina as the sister group to all other cetartiodactylans with 39.7% rootstrap support. However, the solely rooted topology in the AU confidence set is the topology in which the root is placed on the branch leading to Suina (Appendix Table A.6). Similar to Chiroptera and the well-defined clades, the NR-AA model performs better in terms of the BIC score than reversible models in big amino-acid datasets, while the NR-DNA performs worse than reversible models in all datasets (Table A.6). We conclude that neither the nucleotide nor the amino acid data are adequate to confidently infer the root placement of Cetartiodactyla with non-reversible models.

Discussion

In this paper, we introduced a new measure of support for the placement of the root in a phylogenetic tree, the rootstrap support value, and applied it to empirical amino acid and nucleotide datasets inferred using non-reversible models implemented in IQ-TREE (Minh, et al. 2020). The rootstrap is a useful measure because it can be used to assess the statistical support for the placement of the root in any rooted tree, regardless of the rooting method. In a Maximum Likelihood setting, the interpretation of the rootstrap support is similar to the interpretation of the classic nonparametric bootstrap. In a Bayesian setting, the same procedure could be used to calculate the posterior probability of the root placement given a posterior distribution of trees. It is noteworthy that the rootstrap support value is not a measure of the accuracy of the root placement and therefore should not be interpreted as such. However, it provides information about the robustness of the root inference with regard to resampling the data. This interpretation is consistent with the interpretation of the nonparametric bootstrap (Holmes 2003) but with regard to the root placement instead of the whole tree topology.

In addition to the rootstrap support value, we introduced another two metrics; the root branch-length error distance (rBED), and the root split error distance rSED. Similar to the rootstrap metric, these additional metrics can be used with any approach that generates rooted phylogenetic trees. We note that both metrics require the true position of the root to be known (or assumed) and that the rBED requires the rooting method to be able to accurately place the root in a specific position of the root branch.

In this study, we used these and other methods to assess the utility of non-reversible models to root phylogenetic trees in a Maximum Likelihood framework. We focussed on applying these methods to a large and very well curated phylogenomic dataset of mammals, as the mammal phylogeny provides perhaps the best opportunity to find clades for which the root position is known with some confidence. As expected, our results show an exponential increase in the rootstrap support for the true root as we add more information to the MSA. Yet, in some datasets, the rootstrap support drops as more sites are added to the alignment. A careful look into those cases shows that the root inferred by the NR model is in the wrong placement which suggests that the ML inference is inconsistent and therefore unreliable. The inconsistency of the ML could be due to other assumptions (e.g. stationarity or homogeneity) that are severely violated in those datasets.

Our results suggest that non-reversible amino-acid models are more useful for inferring root positions than non-reversible DNA models. One explanation for this difference between the NR-DNA and the NR-AA models is the bigger character-state space of the NR-AA models. These models have 400 parameters (380 rate parameters and 20 amino acid frequencies) whereas NR-DNA models have only 16 parameters (12 rate parameters and 4 nucleotide frequencies). This could allow the NR-AA model to capture the evolutionary process better than the NR-DNA model, potentially providing more information on the root position of the phylogeny. This hypothesis requires some further exploration though, and we note that the actual character-space of amino acids is much smaller than accommodated in NR-DNA models due to functional constraints on protein structure (Dayhoff, et al. 1978).

Another explanation for the difference in performance between the NR-AA and NR-DNA models is that higher compositional heterogeneity in nucleotide datasets may bias tree inference. The fact that each amino acid can be specified by more than one codon, and that synonymous substitutions are more frequent than non-synonymous substitutions, makes amino acid datasets less compositionally heterogeneous than nucleotide datasets. In principle, this bias can be alleviated by removing loci that violate the stationarity and homogeneity assumptions (Naser-Khdour, et al. 2019). Our results suggest that this may be the case for the datasets we analysed: we show that removing loci that violate the stationarity and homogeneity assumptions improves the accuracy and statistical support for the placement of the root. This is not surprising since the robustness of the rootstrap, similar to the bootstrap, relies on the consistency of the inference method, so removing systematic bias should improve its performance.

We used the non-reversible approach to rooting trees along with the rootstrap support to assess the evidence for different root placements in the Chiroptera and Cetartiodactyla. Using the amino acid datasets we found that in both cases, although there tended to be higher rootstrap support for one hypothesis, neither of the current hypotheses for either dataset could be rejected. These results emphasize the importance of the rootstrap support value as a measure of the robustness of the root estimate given the data. In both the Chiroptera and Cetartiodactyla datasets the root placement varied among subsamples of the dataset, and the rootstrap support reflects this uncertainty. However, checking the stability of root placement estimate by randomly subsampling from the whole Chiroptera dataset show an obvious trend towards the

Yinpterochiroptera-Yangochiroptera hypothesis as the dataset increases in size. This trend is consistent with a small number of influential sites or loci having their signal progressively drowned out in favour of the Yinpterochiroptera-Yangochiroptera hypothesis as more data are added to the alignment. In both the Chiroptera and Cetartiodactyla cases, the amino acid data is inadequate to distinguish between certain root placements. On the other hand, in both the Chiroptera and Cetartiodactyla, the nucleotide datasets appear to show stronger support for a single root placement.

Comparing BIC scores of reversible and non-reversible models show that in most of the nucleotide datasets the reversible model was a much better fit to the data than the NR-DNA model. This is likely due to the limitations of the method we used to infer the NR-DNA model. Specifically, when inferring the trees with reversible DNA models, we used a partitioned model such that each partition was able to have an independent DNA substitution model. On the other hand, when we inferred the NR-DNA model we estimated a single model for the entire alignment. Thus, the NR-DNA model we inferred was unable to account for heterogeneity in the evolutionary process among partitions, possibly leading to its worse fit to the data when assessed using BIC scores. This suggests that using either mixture models or partitioned models may improve the fit of non-reversible DNA models to the data. The DNA results are consistent with results from a previous study using the NR-DNA model and RootDigger (Bettisworth and Stamatakis 2020), although that study did not compare the performance of IQ-TREE and RootDigger on empirical datasets. Its results indicate that the NR-DNA model in IQ-TREE could not infer the correct root placement for any of the three tested datasets.

Our results demonstrate that the amino-acid non-reversible model can often be surprisingly accurate for inferring the root placement of phylogenies in the absence of additional information (such as outgroups) or assumptions (such as molecular clocks). In all of the well-defined clades that we examined, the non-reversible amino-acid model successfully identified the root that we identified a-priori as correct, and with very high rootstrap support. Importantly, the non-reversible amino-acid models also tended to fit the data far better than their reversible counterparts did. Indeed, we show that root placements appear to be accurate even with datasets as small as 50 well-curated loci between fairly closely-related taxa such as orders of mammals. Nevertheless, the application of the non-reversible amino acid models to two clades where the root position has previously been contentious failed to shed much additional light on the true root placement. Thus, while we show that the use of non-reversible models certainly has promise, we also show that it is no silver bullet. Yet, as accounting for stationarity and homogeneity would improve the ML inference, using non-reversible models that are also non-stationary and/or non-homogneeous could improve the root placement. However, there is no effective way to do so with the current technology. Some software such as Bio++ (Dutheil and Boussau 2008), BppML (Groussin, et al. 2013), and HAL-HAS (Jayaswal, et al. 2014) relax more than one of the SRH assumptions; but they are rarely used, mainly because of their relatively long run time.

Where a reliable outgroup taxon can be found, without the issues that can confound the inference of root placements using outgroups (Dalevi, et al. 2001; Braun and Kimball 2002; Graham, et al. 2002; Brady, et al. 2006), we suggest relying on the use of outgroups. Nevertheless, where no reliable outgroups can be found, or where there is some reason to question the position of a root inferred using an outgroup (e.g. Bergsten 2005), our study suggests that using non-reversible models can provide a useful additional line of evidence for the position of the root of a phylogeny. We note also that the rootstrap value and the AU test could be used to provide estimates of the uncertainty of root placement using an outgroup taxon

Our work suggests a practical approach to inferring the root of a phylogenetic tree using non-reversible models. First, estimate an unrooted tree topology using the best reversible models available, excluding outgroup sequences. Next, fix the tree topology and use the best non-reversible models available to infer the Maximum Likelihood (ML) root position of that tree. Finally, determine to what extent the ML root position should be trusted. The degree of trust that researchers should put in an inferred ML root position should be influenced by three factors (noting of course that all phylogenetic inferences are susceptible to being misled by model misspecification). First, the fit of the non-reversible model to the data should be better than the fit of the reversible model. This can be assessed using common criteria like AICc or BIC scores. A better fit of the non-reversible model provides some assurance that the data contain sufficient signal that using a non-reversible model is advisable in the first place. Our results show that the root placement was inferred correctly with high rootstrap support in 12 out of the 13 datasets in which the non-reversible model was preferable. In the absence of a better fit for a non-reversible model, we do not think any inferred ML root position should be trusted. Second, root positions with higher rootstrap support should be trusted more, because a higher rootstrap support indicates less variance among sites in the signal for the placement of the root. Third, ML root positions should be trusted more when the number of root placements included in the confidence set of an AU test is small because a smaller confidence set indicates that there is less uncertainty in the root placement when the analysis is conditioned on the full alignment and the unrooted ML tree topology. A conservative approach to inferring root placements with non-reversible models would be to consider any root placement that has a substantial fraction of the rootstrap support and/or is included in the set of possible root placements identified by the AU test as a possible root placement given the assumptions of the model.

We hope that the combination of non-reversible models, rootstrap support, and AU tests will add another tool to the phylogeneticist’s arsenal when it comes to inferring rooted phylogenies.

Funding

This work was supported by an Australian Research Council (Grant No. DP200103151 to R.L., B.Q.M.) and by a Chan-Zuckerberg Initiative grant to B.Q.M and R.L.

References

Agnarsson I, Zambrana-Torrelio CM, Flores-Saldana NP, May-Collado LJ. 2011. A time-calibrated species-level phylogeny of bats (Chiroptera, Mammalia). PLoS Curr 3:RRN1212.

Allman ES, Degnan JH, Rhodes JA. 2011. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. J. Math. Biol. 62:833-862.

Baldauf SL, Palmer JD. 1993. Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. Proceedings of the National Academy of Sciences 90:11558-11562.

Bergsten J. 2005. A review of long-branch attraction. Cladistics 21:163-193.

Bettisworth B, Stamatakis A. 2020. RootDigger: a root placement program for phylogenetic trees. bioRxiv:2020.2002.2013.935304.

Boussau B, Szollosi GJ, Duret L, Gouy M, Tannier E, Daubin V. 2013. Genome-scale coestimation of species and gene trees. Genome Res. 23:323-330.

Brady SG, Schultz TR, Fisher BL, Ward PS. 2006. Evaluating alternative hypotheses for the early evolution and diversification of ants. Proc Natl Acad Sci U S A 103:18172-18177.

Braun EL, Kimball RT. 2002. Examining Basal avian divergences with mitochondrial sequences: model complexity, taxon sampling, and sequence length. Syst. Biol. 51:614-625.

Cherlin S, Heaps SE, Nye TMW, Boys RJ, Williams TA, Embley TM. 2018. The Effect of Nonreversibility on Inferring Rooted Phylogenies. Mol. Biol. Evol. 35:984-1002.

Chernomor O, von Haeseler A, Minh BQ. 2016. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. Syst. Biol. 65:997-1008.

Dalevi D, Hugenholtz P, Blackall LL. 2001. A multiple-outgroup approach to resolving division-level phylogenetic relationships using 16S rDNA data. Int. J. Syst. Evol. Microbiol. 51:385-391.

Dayhoff M. 1987. A model of evolutionary change in proteins. Atlas of protein sequence and structure 5:suppl. 3.

Dayhoff M, Schwartz R. 1980. Prokaryote evolution and the symbiotic origin of eukaryotes. Endocytobiology: endosymbiosis and cell biology: a synthesis of recent research 1:63-84.

Dayhoff M, Schwartz R, Orcutt B. 1978. A model of evolutionary change in proteins. Atlas of protein sequence and structure 5:345-352.

Drummond AJ, Ho SY, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Duchene DA, Tong KJ, Foster CSP, Duchene S, Lanfear R, Ho SYW. 2020. Linking Branch Lengths across Sets of Loci Provides the Highest Statistical Support for Phylogenetic Inference. Mol. Biol. Evol. 37:1202-1210.

Farris JS. 1972. Estimating Phylogenetic Trees from Distance Matrices. Am. Nat. 106:645-&.

Gatesy J, Springer MS. 2017. Phylogenomic red flags: Homology errors and zombie lineages in the evolutionary diversification of placental mammals. Proc Natl Acad Sci U S A 114:E9431-E9432.

Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, Bowman BJ, Manolson MF, Poole RJ, Date T, Oshima T, et al. 1989. Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. Proc Natl Acad Sci U S A 86:6661-6665.

Graham SW, Olmstead RG, Barrett SC. 2002. Rooting phylogenetic trees with distant outgroups: a case study from the commelinoid monocots. Mol. Biol. Evol. 19:1769-1781.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160-174.

Heaps SE, Nye TM, Boys RJ, Williams TA, Embley TM. 2014. Bayesian modelling of compositional heterogeneity in molecular phylogenetics. Stat Appl Genet Mol Biol 13:589-609.

Holmes S. 2003. Bootstrapping phylogenetic trees: Theory and methods. Statistical Science 18:241-255.

Huelsenbeck JP, Bollback JP, Levine AM. 2002. Inferring the Root of a Phylogenetic Tree. Syst. Biol. 51:32-43.

Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci U S A 86:9355-9359.

Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW, Leebens-Mack J, Muller KF, Guisinger-Bellian M, Haberle RC, Hansen AK, et al. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci U S A 104:19369-19374.

Jayaswal V, Ababneh F, Jermiin LS, Robinson J. 2011. Reducing model complexity of the general Markov model of evolution. Mol. Biol. Evol. 28:3045-3059.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8:275-282.

Jukes TH, Cantor C. 1969. Evolution of protein molecules. In: Munro HN, editor. In Mammalian Protein Metabolism.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14:587-589.

Kimura M. 1980. A Simple Method for Estimating Evolutionary Rates of Base Substitutions through Comparative Studies of Nucleotide-Sequences. J. Mol. Evol. 16:111-120.

Kocot KM, Cannon JT, Todt C, Citarella MR, Kohn AB, Meyer A, Santos SR, Schander C, Moroz LL, Lieb B, et al. 2011. Phylogenomics reveals deep molluscan relationships. Nature 477:452-456.

Lake JA, Herbold CW, Rivera MC, Servin JA, Skophammer RG. 2007. Rooting the tree of life using nonubiquitous genes. Mol. Biol. Evol. 24:130-136.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25:1307-1320.

Liu L, Zhang J, Rheindt FE, Lei F, Qu Y, Wang Y, Zhang Y, Sullivan C, Nie W, Wang J, et al. 2017. Genomic evidence reveals a radiation of placental mammals uninterrupted by the KPg boundary. Proc Natl Acad Sci U S A 114:E7282-E7290.

Lyons-Weiler J, Hoelzer GA, Tausch RJ. 1998. Optimal outgroup analysis. Biol. J. Linn. Soc. 64:493-511.

Maddison WP, Donoghue MJ, Maddison DR. 1984. Outgroup Analysis and Parsimony. Systematic Zoology 33:83-103.

Mai U, Sayyari E, Mirarab S. 2017. Minimum variance rooting of phylogenetic trees and implications for species tree reconstruction. PLoS One 12:e0182238.

Meganathan PR, Pagan HJ, McCulloch ES, Stevens RD, Ray DA. 2012. Complete mitochondrial genome sequences of three bats species and whole genome mitochondrial analyses reveal patterns of codon bias and lend support to a basal split in Chiroptera. Gene 492:121-129.

Meredith RW, Janecka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simao TL, Stadler T, et al. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. Science 334:521-524.

Milinkovitch MC, Lyons-Weiler J. 1998. Finding optimal ingroup topologies and convexities when the choice of outgroups is not obvious. Mol. Phylogen. Evol. 9:348-357.

Minh BQ, Hahn MW, Lanfear R. 2020. New methods to calculate concordance factors for phylogenomic datasets. Mol. Biol. Evol.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. 37:1530-1534.

Moore MJ, Bell CD, Soltis PS, Soltis DE. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc Natl Acad Sci U S A 104:19363-19368.

Nardi F, Spinsanti G, Boore JL, Carapelli A, Dallai R, Frati F. 2003. Hexapod origins: monophyletic or paraphyletic? Science 299:1887-1889.

Naser-Khdour S, Minh BQ, Zhang W, Stone EA, Lanfear R. 2019. The Prevalence and Impact of Model Violations in Phylogenetic Analysis. Genome Biol Evol.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DT, Manuel M, Worheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.

Ren M, Sun HJ, Bo SQ, Zhang SY, Hua PY. 2018. Parallel amino acid deletions of prestin protein in two dramatically divergent bat lineages suggest the complexity of the evolution of echolocation in bats. Acta Chiropterologica 20:311-317.

Reyes-Amaya N, Flores D. 2019. Hypophysis size evolution in Chiroptera. Acta Chiropterologica 21:65-74.

Rivera MC, Lake JA. 1992. Evidence That Eukaryotes and Eocyte Prokaryotes Are Immediate Relatives. Science 257:74-76.

Shen XX, Hittinger CT, Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat Ecol Evol 1:126.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492-508.

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114-1116.

Smith AB. 1994. Rooting Molecular Trees - Problems and Strategies. Biol. J. Linn. Soc. 51:279-292.

Squartini F, Arndt PF. 2008. Quantifying the Stationarity and Time Reversibility of the Nucleotide Substitution Process. Molecular Biology and Evolution 25:2525-2535.

Steenkamp ET, Wright J, Baldauf SL. 2006. The protistan origins of animals and fungi. Mol. Biol. Evol. 23:93-106.

Swofford D, Olsen G, Waddell P. 1996. Phylogenetic inference. In: David M. Hillis CM, Barbara K. Mable, editor. Molecular Systematics, 2nd edn: Sunderland, Mass. : Sinauer Associates. p. 407-513.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10:512-526.

Tavaré S. 1986. Some probabilistic and statistical probles in the analysis of DNA sequences. Lectures on Mathematics in the Life Sciences 17.

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

Tsagkogeorga G, Parker J, Stupka E, Cotton JA, Rossiter SJ. 2013. Phylogenomic analyses elucidate the evolutionary relationships of bats. Curr. Biol. 23:2262-2267.

Watrous LE, Wheeler QD. 1981. The out-Group Comparison Method of Character Analysis. Systematic Zoology 30:1-11.

Whelan NV, Kocot KM, Moroz LL, Halanych KM. 2015. Error, signal, and the placement of Ctenophora sister to all other animals. Proc Natl Acad Sci U S A 112:5773-5778.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18:691-699.

Williams KP, Gillespie JJ, Sobral BW, Nordberg EK, Snyder EE, Shallom JM, Dickerman AW. 2010. Phylogeny of gammaproteobacteria. J. Bacteriol. 192:2305-2314.

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

Wu S, Edwards S, Liu L. 2019. Data from: Genome-scale DNA sequence data and the evolutionary history of placental mammals. In: Figshare.

Wu S, Edwards S, Liu L. 2018. Genome-scale DNA sequence data and the evolutionary history of placental mammals. Data Brief 18:1972-1975.

Yang ZH, Roberts D. 1995. On the Use of Nucleic-Acid Sequences to Infer Early Branchings in the Tree of Life. Mol. Biol. Evol. 12:451-458.

Yap VB, Speed T. 2005. Rooting a phylogenetic tree with nonreversible substitution models. BMC Evol. Biol. 5:2.

Yu Y, Warnow T, Nakhleh L. 2011. Algorithms for MDC-based multi-locus phylogeny inference: beyond rooted binary gene trees on single alleles. J. Comput. Biol. 18:1543-1559.

Zhang SQ, Che LH, Li Y, Dan L, Pang H, Slipinski A, Zhang P. 2018. Evolutionary history of Coleoptera revealed by extensive sampling of genes and species. Nat Commun 9:205.

Zhou X, Xu S, Yang Y, Zhou K, Yang G. 2011. Phylogenomic analyses and improved resolution of Cetartiodactyla. Mol Phylogenet Evol 61:255-264.

Appendix

Algorithm A.1.

Since we want to define the well-defined clades for the non-reversible analysis based on the bootstrap support values and the concordance factors values we will use the following algorithm to find well-defined clades in the whole MSA:

(1) First, estimate the unrooted topology and the bootstrap support values using the ultrafast bootstrap (UFBoot) with 1000 replicates (Hoang, et al. 2018)

```
iqtree2 –s ALIGNMENT_FILE –p PARTITION_FILE –B 1000 --prefix REV
```

Where ALIGNMENT_FILE is the MSA file, PARTITION_FILE is the partition file (it can be the same as the MSA file), -p is the option for edge-linked substitution rates,–B is option for UFBoot with 1000 replicates, and --prefix so all the output files will be named REV.*.

(2) Infer the single-locus trees

```
iqtree2 –s ALIGNMENT_FILE –S PARTITION_FILE --prefix LOCI
```

Like –p option, the –S option performs model selection for each loci separately. However, unlike –p option, the –S option infer a separate tree for each loci. All output files are of the form LOCI.*.

(3) calculate the gCF and sCF (Minh, et al. 2018) for every branch of the ML species tree

```
Iqtree2 –t REV.treefile --gcf LOCI.treefile –s ALIGNMENT_FILE --scf 100 --prefix CONCORD
```

Where REV.treefile is the unrooted species tree, and LOCI.treefile is file with all the unrooted loci trees.

Algorithm A.2.

For each sub-dataset that we want to infer the root placement using the non-reversible model. However, as the non-reversible model is computationally expensive, we first find the best partitioning scheme using the reversible model and then use this scheme to find the non-reversible model parameters and infer the rooted tree.

(1) Find the best partitioning scheme using the best-fit reversible model

```
    iqtree2 –s ALIGNMENT_FILE –p PARTITION_FILE --prefix REV
```

(2) Estimate the non-reversible models' parameters using the best partitioning scheme from the previous step and unrooted tree as the initial tree

For NR-AA:

```
iqtree2 –s ALIGNMENT_FILE –p REV.best_scheme.nex –t REV.treefile

--model-joint NONREV –B 1000 --prefix AA_NONREV
```

For NR-DNA:

```
iqtree2 –s ALIGNMENT_FILE –p REV.best_scheme.nex –t REV.treefile

--model-joint 12.12 –B 1000 --prefix DNA_NONREV
```

Where NONREV and 12.12 are the substitution models for NR-AA and NR-DNA as defined in IQ-TREE, respectively.

(3) Compare the BIC scores of the reversible and non-reversible models.

If $BIC_{NONREV} > BIC_{REV}$, then using non-reversible models to infer the rooted topology is not advised due to over-parameterization of the data. In this case, an alternative rooting method is recommended if possible.

Algorithm A.3.

We apply the AU test (Shimodaira 2002) to all the possible rooting placements using the following command-line:

For NR-AA:

```
iqtree2 –s ALIGNMENT_FILE –p REV.best_scheme.nex –model-joint NONREV
--root-test –zb 1000 –au –te NONREV.treefile --prefix TOP
```

For NR-DNA:

```
iqtree2 –s ALIGNMENT_FILE –p REV.best_scheme.nex –model-joint 12.12 -
-root-test –zb 1000 –au –te NONREV.treefile --prefix TOP
```

Where --root-test will re-root the tree on every branch, -zb option for specifying the number of RELL replicates (Kishino, et al. 1990), and -au option to perform the AU test.

TABLE A.1. Last decade's relevant literature of the well-defined clades that we used in this study for estimating the root.

Clade	Study Reference	Root placement
Afrotheria	(Elliot and Crespi 2009)	{Afroinsectiphilia, Paenungulata}
	(Asher, et al. 2009)	{Afroinsectiphilia, Paenungulata}
	(Poulakakis and Stamatakis 2010)	{Afroinsectiphilia, Paenungulata}
	(Phillips and Penny 2010)	{Afroinsectiphilia, Paenungulata}
	(Romiguier, et al. 2010)	{Afroinsectiphilia, Paenungulata}
	(Kuntner, et al. 2011)	{Afroinsectiphilia, Paenungulata}
	(Meredith, et al. 2011)	{Afroinsectiphilia, Paenungulata}
	(Svartman and Stanyon 2012)	{Afroinsectiphilia, Paenungulata}
	(Lartillot and Delsuc 2012)	{Afroinsectiphilia, Paenungulata}
	(dos Reis, et al. 2012)	{Afroinsectiphilia, Paenungulata}
	(Benoit, et al. 2013)	{Afroinsectiphilia, Paenungulata}
	(Wu, et al. 2014)	{Afroinsectiphilia, Paenungulata}
	(Gheerbrant, et al. 2014)	{Afroinsectiphilia, Paenungulata}
	(Halliday, et al. 2015)	{Afroinsectiphilia, Paenungulata}
	(Puttick and Thomas 2015)	{Afrosoricida },{Afroinsectiphilia, Paenungulata}
	(Foley, et al. 2016)	{Afroinsectiphilia, Paenungulata}
	(Liu, et al. 2017)	{Afroinsectiphilia, Paenungulata}
	(Wu, et al. 2017)	{Afroinsectiphilia, Paenungulata}
Primates	(Fabre, et al. 2009)	{Strepsirrhini, Haplorrhini}
	(Chatterjee, et al. 2009)	{Strepsirrhini, Haplorrhini}
	(Matsui, et al. 2009)	{Trasiiformes},{Strepsirrhini, Haplorrhini}
	(Romiguier, et al. 2010)	{Strepsirrhini, Haplorrhini}
	(Perelman, et al. 2011)	{Strepsirrhini, Haplorrhini}
	(Meredith, et al. 2011)	{Strepsirrhini, Haplorrhini}
	(Jameson, et al. 2011)	{Strepsirrhini, Haplorrhini}
	(Diogo and Wood 2011)	{Strepsirrhini, Haplorrhini}
	(Springer, et al. 2012)	{Strepsirrhini, Haplorrhini}
	(dos Reis, et al. 2012)	{Strepsirrhini, Haplorrhini}
	(Steiper and Seiffert 2012)	{Strepsirrhini, Haplorrhini}
	(Lartillot and Delsuc 2012)	{Strepsirrhini, Haplorrhini}
	(Finstermeier, et al. 2013)	{Strepsirrhini, Haplorrhini}
	(Hartig, et al. 2013)	{Strepsirrhini, Haplorrhini}
	(Kumar, et al. 2013)	{Strepsirrhini, Haplorrhini}
	(Pozzi, et al. 2014)	{Strepsirrhini, Haplorrhini}
	(Wu, et al. 2014)	{Strepsirrhini, Haplorrhini}
	(Pattinson, et al. 2015)	{Strepsirrhini, Haplorrhini}
	(Kari, et al. 2015)	{Strepsirrhini, Haplorrhini}
	(Herrera and Davalos 2016)	{Strepsirrhini, Haplorrhini}

	(Liu, et al. 2017)	{Strepsirrhini, Haplorrhini}
	(Wu, et al. 2017)	{Strepsirrhini, Haplorrhini}
	(Monson and Hlusko 2018)	{Strepsirrhini, Haplorrhini}
	(Reis, et al. 2018)	{Strepsirrhini, Haplorrhini}
	(Zhang, et al. 2019)	{Strepsirrhini, Haplorrhini}
Myomorpha	(Blanga-Kanfi, et al. 2009)	{Muroidea, Dipodoidea}
	(Churakov, et al. 2010)	{Muroidea, Dipodoidea}
	(Hao, et al. 2011)	{Muroidea, Dipodoidea}
	(Meredith, et al. 2011)	{Muroidea, Dipodoidea}
	(Horn, et al. 2011)	{Muroidea, Dipodoidea}
	(Fabre, et al. 2012)	{Muroidea, Dipodoidea}
	(Wu, et al. 2012)	{Muroidea, Dipodoidea}
	(Schenk, et al. 2013)	{Muroidea, Dipodoidea}
	(Wu, et al. 2014)	{Muroidea, Dipodoidea}
	(Yue, et al. 2015)	{Muroidea, Dipodoidea}
	(Liu, et al. 2017)	{Muroidea, Dipodoidea}
	(Wu, et al. 2017)	{Muroidea, Dipodoidea}
	(Tavares and Seuanez 2018)	{Muroidea, Dipodoidea}
	(Swanson, et al. 2019)	{Muroidea, Dipodoidea}
	(Hedrick, et al. 2020)	{Muroidea, Dipodoidea}
Carnivora	(Finarelli and Flynn 2009)	{Feliformia,Caniformia}
	(Agnarsson, et al. 2010)	{Feliformia,Caniformia}
	(Eizirik, et al. 2010)	{Feliformia,Caniformia}
	(Stankowich, et al. 2011)	{Feliformia,Caniformia}
	(Nyakatura and Bininda-Emonds 2012)	{Feliformia,Caniformia}
	(Lartillot and Delsuc 2012)	{Feliformia,Caniformia}
	(dos Reis, et al. 2012)	{Feliformia,Caniformia}
	(Wu, et al. 2014)	{Feliformia,Caniformia}
	(Tomiya and Tseng 2016)	{Feliformia,Caniformia}
	(Panciroli, et al. 2017)	{Feliformia,Caniformia}
	(Liu, et al. 2017)	{Feliformia,Caniformia}
	(Wu, et al. 2017)	{Feliformia,Caniformia}
	(Polly, et al. 2017)	{Feliformia,Caniformia}
	(Machado, et al. 2018)	{Feliformia,Caniformia}
Bovidae	(Bibi, et al. 2009)	{Bovinae}
	(Hassanin, et al. 2012)	{Bovinae}
	(Yang, et al. 2013)	{Bovinae}
	(Wu, et al. 2014)	{Bovinae}
	(Bibi 2013)	{Bovinae}
	(Liu, et al. 2017)	{Bovinae}
	(Wu, et al. 2017)	{Bovinae}
	(Chen, et al. 2019)	{Bovinae}

TABLE A.2. The well-defined clades that we used in this study for estimating the root.

Clade	No. taxa	Dataset	branch	%BS	gCF	gDF1	gDF2	sCF
Afrotheria	7	AA	leading	100.0	55.18	0.32	0.66	55.34
			descen.1	100.0	24.68	5.38	8.30	35.64
			descen.2	100.0	55.43	1.41	1.01	55.12
		DNA	leading	100.0	52.72	0.18	0.42	50.56
			descen.1	100.0	17.97	6.23	8.07	32.87
			descen.2	100.0	50.95	0.96	1.41	51.22
Primates	16	AA	leading	100.0	25.10	2.14	2.67	41.22
			descen.1	100.0	25.77	8.82	7.91	37.91
			descen.2	100.0	45.10	1.25	1.75	55.00
		DNA	leading	100.0	28.04	2.19	2.79	38.79
			descen.1	100.0	27.29	9.29	7.63	35.82
			descen.2	100.0	47.64	1.42	1.67	50.39
Myomorpha	7	AA	leading	100.0	58.10	5.94	5.9	48.09
			descen.1	100.0	84.47	0.79	0.61	73.88
		DNA	leading	100.0	57.47	5.36	5.74	43.32
			descen.1	100.0	83.95	0.62	0.50	65.66
Carnivora	9	AA	leading	100.0	60.22	1.01	1.51	64.41
			descen.1	100.0	82.06	0.53	0.76	94.15
			descen.2	100.0	49.80	5.62	7.87	50.53
		DNA	leading	100.0	53.63	6.11	6.51	45.8
			descen.1	100.0	84.23	0.50	0.62	92.34
			descen.2	100.0	51.18	3.88	6.2	50.73
Bovidae	5	AA	leading	100.0	85.66	0.18	0.16	89.47
			descen.1	100.0	81.28	3.9	2.38	93.25
			descen.2	100.0	71.83	5.71	5.28	79.17
		DNA	leading	100.0	85.53	0.28	0.21	85.37
			descen.1	100.0	81.78	3.64	2.27	90.19
			descen.2	100.0	72.21	5.38	4.97	70.74

Note: The bootstrap, gCF, gDF1, gDF2, and sCF values are for the branch leading to that clade and the first direct descendants of the clade, respectively.

Table A.3. Rootstrap support value, number of alternative root placements in the AU test confidence set, rBED, and rSED for the amino-acid rooted trees of each clade.

Clade	#loci	#sites	%RS support for true root	Is true root in CS?	root placements in CS	rBED	rSED
Carnivora	5,162	3,050,199	100%	Yes	0	0.0000 – 0.0273	0
	516	295,962	99.8%	Yes	0	0.0000 – 0.0274	0
	52	33,404	99.1%	Yes	0	0.0000 – 0.0242	0
	5	2,789	4.6%	No	1	0.0074 – 0.0740	1
Bovidae	5,162	3,050,199	100%	Yes	0	0.0000 – 0.0128	0
	516	310,353	100%	Yes	0	0.0000 – 0.0130	0
	52	28,729	99.9%	Yes	0	0.0000 – 0.0092	0
	5	3,222	41.9%	Yes	0	0.0000 – 0.0210	0
Myomorpha	5,162	3,050,199	100.0%	Yes	0	0.0000 – 0.0842	0
	516	309,459	100.0%	Yes	0	0.0000 – 0.0886	0
	52	35,494	100.0%	Yes	0	0.0000 – 0.0679	0
	5	4,438	36.6%	Yes	0	0.0000 – 0.1135	0
Primates	5,162	3,050,199	100.0%	Yes	0	0.0000 – 0.0092	0
	516	302,781	87.9%	Yes	0	0.0000 – 0.0090	0
	52	30,225	59.8%	Yes	0	0.0000 – 0.0106	0
	5	4,426	66.1%	Yes	0	0.0000 – 0.0076	0
Afrotheria	5,162	3,050,199	100.0%	Yes	0	0.0000 – 0.0077	0
	516	292,632	98.5%	Yes	0	0.0000 – 0.0066	0
	52	28,904	29.2%	No	1	0.0004 – 0.0127	1
	5	3,265	62.1%	Yes	0	0.0000 – 0.0120	0

Table A.4. Rootstrap support value, number of alternative root placements in the AU test confidence set, rBED, and rSED for the nucleotide rooted trees of each clade.

Clade	#loci	#sites	%RS support for true root	Is true root in CS?	root placements in CS	rBED	rSED
Carnivora	15,486	9,150,597	100%	Yes	0	0.0000 – 0.0229	0
	1,548	917,265	100%	Yes	1	0.0000 – 0.0273	0
	154	99,334	99.1%	Yes	0	0.0000 – 0.0238	0
	15	11,197	17.5%	No	1	0.0199 – 0.0420	1
Bovidae	15,486	9,150,597	100.0%	Yes	0	0.0000 – 0.0269	0
	1,548	910,022	100.0%	Yes	0	0.0000 – 0.0269	0
	154	91,674	40.8%	Yes	0	0.0000 – 0.0196	0
	15	5,560	1.4%	No	2	0.0036 – 0.0198	2
Myomorpha	15,485	9,149,793	73.2%	Yes	1	0.0000 – 0.1292	0
	1,548	898,643	25.9%	Yes	1	0.0048 – 0.0150	1
	154	87,033	5.4%	Yes	1	0.0159 – 0.0253	1
	15	9,433	15.7%	No	1	0.0352 – 0.0431	2
Primates	15,486	9,150,597	99.7%	Yes	0	0.0000 – 0.0096	0
	1,548	912,110	87.0%	Yes	0	0.0000 – 0.0087	0
	154	93,830	31.1%	Yes	0	0.0132 – 0.0300	1
	15	8,242	3.0%	Yes	1	0.0176 – 0.0221	4
Afrotheria	15,486	9,150,597	0.0%	No	1	0.0157 – 0.0268	2
	1,548	924,981	21.0%	No	1	0.0133 – 0.0251	2
	154	92,545	44.4%	No	1	0.0000 – 0.0073	0
	15	6,697	4.7%	No	1	0.0268 – 0.0428	2

Table A.5. Number of sites, root placement, RS support for the ML root placement, better BIC score (between R and NR models), and root placements in the AU confidence set for Chiroptera.

	dataset	#sites	Root placement	RS% ML root	Better BIC	AU CS
DNA-NR	Whole dataset	9,149,793	Y-Y	100.0%	R	{Y-Y}
	Subsampled 10%	921,910	Y-Y	99.9%	R	{Y-Y}
	Subsampled 1%	85,437	Y-Y	41.6%	R	{Y-Y}
	Subsampled 0.1%	7,804	Y-Y	9.6%	R	{Y-Y}
	MaxSym test	6,854,459	Y-Y	90.1%	NR	{Y-Y}
	Highest GLS	8,769,211	Y-Y	100.0%	R	{Y-Y}
AA-NR	Whole dataset	3,050,199	Y-Y	65.5%	NR	{Y-Y, M-M, R, P}
	Subsampled 10%	309,745	Y-Y	51.0%	NR	{Y-Y, M-M}
	Subsampled 1%	33,365	R	38.1%	R	{Y-Y, R, P}
	Subsampled 0.1%	3,955	M-M	70.2%	R	{Y-Y, M-M}
	MaxSym test	2,934,731	Y-Y	65.9%	NR	{Y-Y, M-M, R }
	Highest GLS	3,027,379	Y-Y	76.6%	NR	{Y-Y, M-M}

Note: Y-Y refers to Yinptero-Yango hypothesis, M-M refers to Micro-Mega hypothesis, R refers to Rhinolophoidea and P refers to Pteropus (Flying foxes). Highest GLS refers to dataset where the top 1% loci with the highest ΔGLS removed.

Table A.6. Number of sites, root placement, RS support for the ML root placement, better BIC score (between R and NR models, and root placements in the AU confidence set for Cetartiodactyla.

	dataset	#sites	Root placement	RS% ML root	Better AICc	AU CS
DNA-NR	Whole dataset	9,149,793	T	71.0%	R	{T}
	Subsampled 10%	915,274	T	84.5%	R	{T}
	Subsampled 1%	90,371	T+S	38.9%	R	{S}
	Subsampled 0.1%	6,003	T	41.6%	R	{T}
	MaxSym test	7,565,809	T	68.7%	R	{T}
	Highest GLS	8,557,372	T+S	39.7%	R	{S}
AA-NR	Whole dataset	3,050,199	T+S	71.8%	NR	{T+S, T, S}
	Subsampled 10%	293,123	T	62.9%	NR	{T+S}
	Subsampled 1%	31,253	T+S	50.2%	R	{T+S, T}
	Subsampled 0.1%	1,480	B	40.7%	R	{T+S, T}
	MaxSym test	2,997,304	T+S	63.6%	NR	{T+S, T}
	Highest GLS	3,024,886	T+S	63.2%	NR	{T+S, T}

Note: T refers to Tylopoda, S refers to Suina, and B refers to Bovidae. Highest GLS refers to dataset where the top 1% loci with the highest ΔGLS removed.

Table A.7. Likelihood, number of free parameters, and BIC score of the amino-acid reversible and non-reversible models.

		Rev-AA			NR-AA		
Clade	#loci	logL	#FP	BIC	logL	#FP	BIC
Afrotheria	5,162	-14330208.59	21172	28976530	-14297752.78	9716	28740572
	516	-1357177.681	2117	2741001	-1354652.159	1320	2725919
	52	-136982.9062	220	276226	-136452.7464	487	277908
	5	-19001.63	58	38473	-18665.2888	401	40575
Bovidae	5,162	-9495698.792	16456	19237097	-9526590.714	6026	19143154
	516	-969922.7367	1716	1961545	-972933.315	957	1957968
	52	-90113.4964	100	181254	-90307.5227	443	185163
	5	-10558.2498	11	21205	-10429.3843	392	24025
Carnivora	5,162	-11771808.39	19563	23835706	-11783521.78	8525	23694328
	516	-1159792.644	1907	2343610	-1160442.419	1205	2336065
	52	-133619.2048	288	270238	-133736.2131	479	272462
	5	-12401.6075	37	25097	-12212.3584	399	27590
Myomorpha	5,162	-13717877.69	21337	27754332	-13704191.03	9710	27553359
	516	-1410118.562	2041	2846041	-1407094.056	1320	2830876
	52	-159843.2064	219	321981	-159218.9652	486	323530
	5	-19717.5842	38	39754	-19553.1957	400	42466
Primates	5,162	-14060075.69	19597	28412749	-14035582.5	9433	28212006
	516	-1409968.454	1965	2844737	-1406902.357	1320	2830464
	52	-135178.7103	209	272514	-134377.0976	495	273861
	5	-18618.1041	74	37857	-18440.361	417	40382

Table A.8. Likelihood, number of free parameters, and BIC score of nucleotide reversible and non-reversible models.

		Rev-DNA			NR-DNA		
Clade	#loci	logL	#FP	BIC	logL	#FP	BIC
Afrotheria	15,486	-25649218.58	83714	52640316	-26103942.35	26592	52634137
	1,548	-2592184.776	8317	5298625	-2638008.645	2681	5312848
	154	-269601.7756	822	548603	-274501.0773	290	552318
	15	-19264.4287	71	39154	-19465.3026	50	39371
Bovidae	15,486	-14276063.64	60132	29516003	-14675053.78	16919	29621308
	1,548	-1419112.739	5975	2920210	-1458435.435	1711	2940348
	154	-144047.1052	562	294516	-147514.636	185	297143
	15	-8365.8073	64	17284	-8579.0994	35	17460
Carnivora	15,485	-20189220.63	73591	41558056	-20634927.77	22881	41636623
	1,548	-2042901.629	7358	4186822	-2088967.935	2306	4209595
	154	-223544.4981	766	455903	-228291.2125	252	459482
	15	-20813.3401	72	42298	-21110.0594	42	42612
Myomorpha	15,486	-25576581.35	82612	52477370	-25988907.4	26527	52403022
	1,548	-2479841.109	8281	5073203	-2521869.523	2692	5080643
	154	-232786.927	841	475139	-236472.9178	284	476176
	15	-25439.8935	97	51768	-25802.2226	49	52053
Primates	15,486	-25127036.02	79361	51526169	-25556510.21	25658	51524299
	1,548	-2531565.112	8046	5173550	-2575829.296	2631	5187765
	154	-256432.8736	841	522495	-261055.6783	306	525615
	15	-17934.338	109	36852	-18205.7104	65	36998

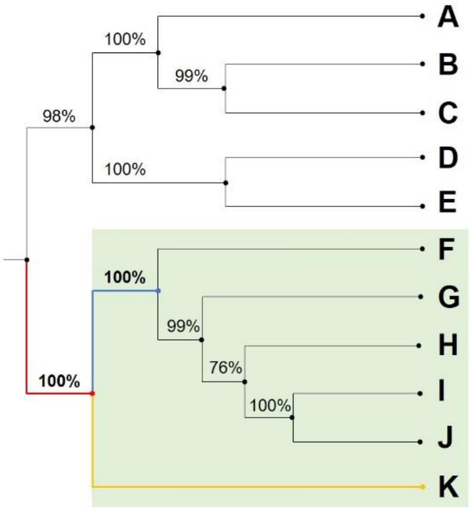


Figure A.1. ML tree with the bootstrap support values for each branch. The only clade that contains at least five taxa and has 100% bootstrap support at the branch leading to that clade and at the first direct descendants in the clade is the green tree (F-K). The red branch is the branch leading to the clade, the blue and the yellow branches are the descendant branches. Note, since the yellow branch is a tip, the bootstrap support is 100%.

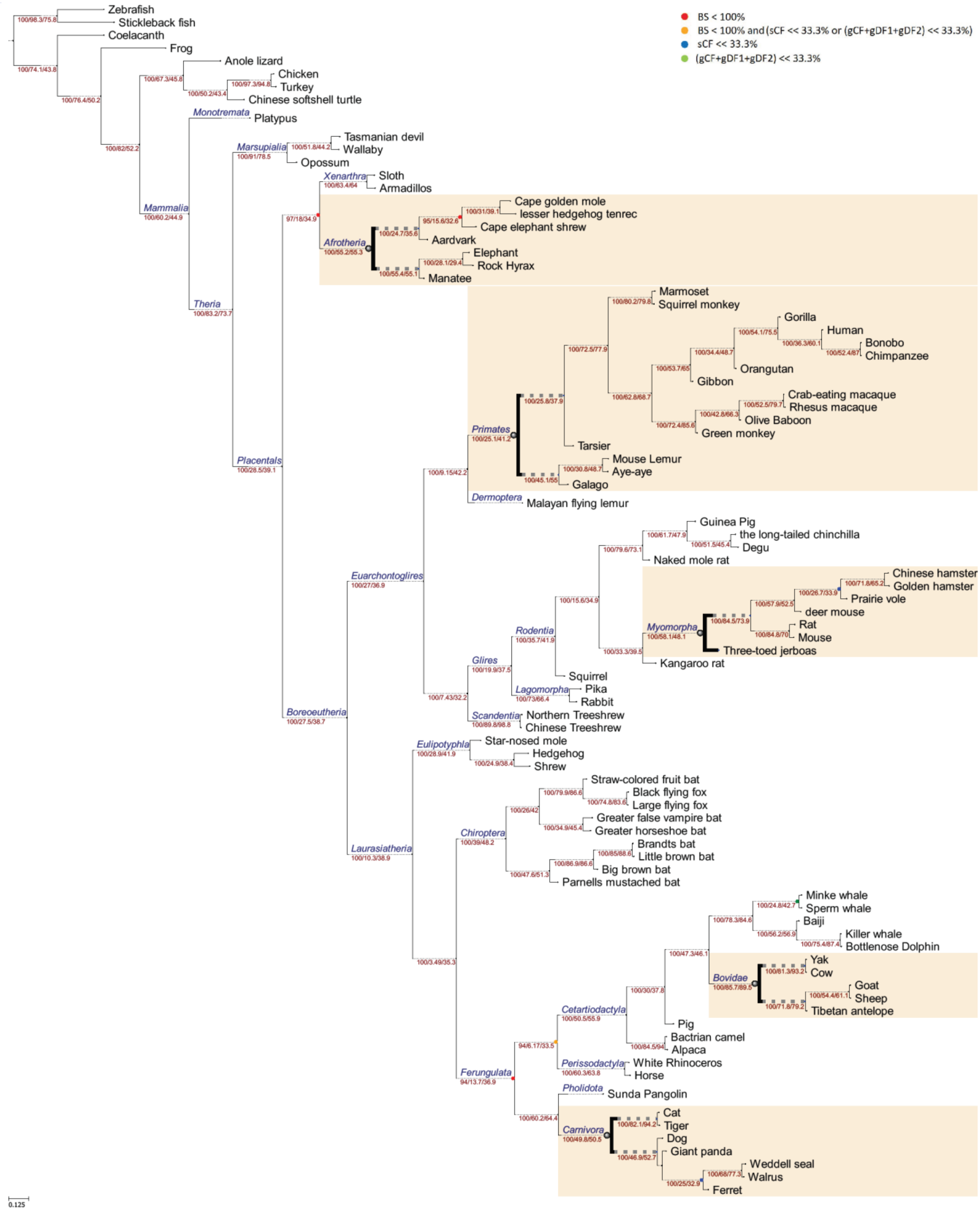


Figure A.2. The ML tree inferred from the whole concatenated AA alignment from (Wu, et al. 2018) and rooted on non-mammalian outgroup taxa. Bold branches present the well-defined clades we use in this study.

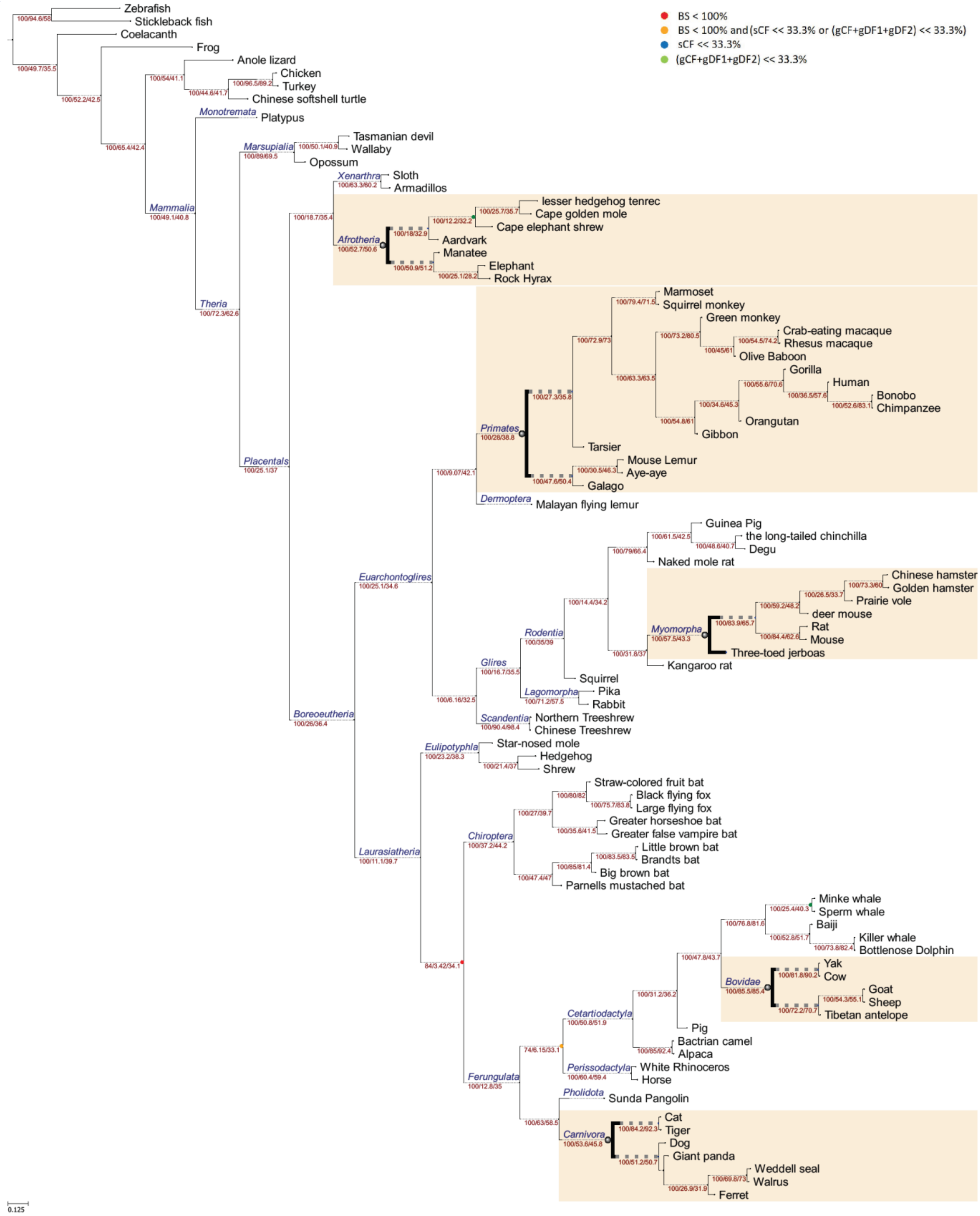


Figure A.3. The ML tree inferred from the whole concatenated DNA alignment from (Wu, et al. 2018) and rooted on non-mammalian outgroup taxa.

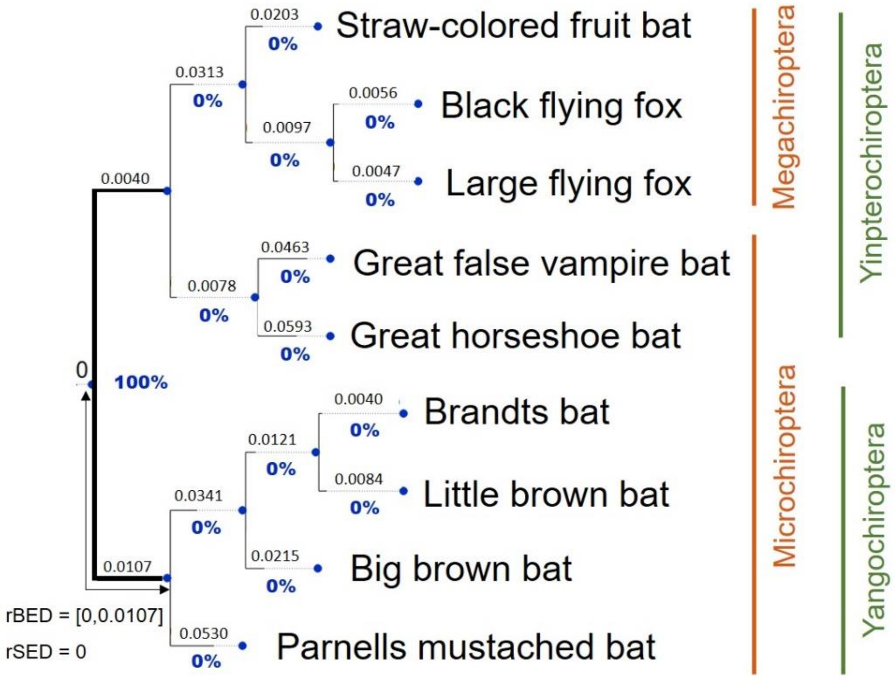


Figure A.4. The ML rooted tree of as inferred from the whole Chiroptera nucleotide dataset. Bold branches are branches in the AU confidence set. Blue values under each branch are the rootstrap support values.

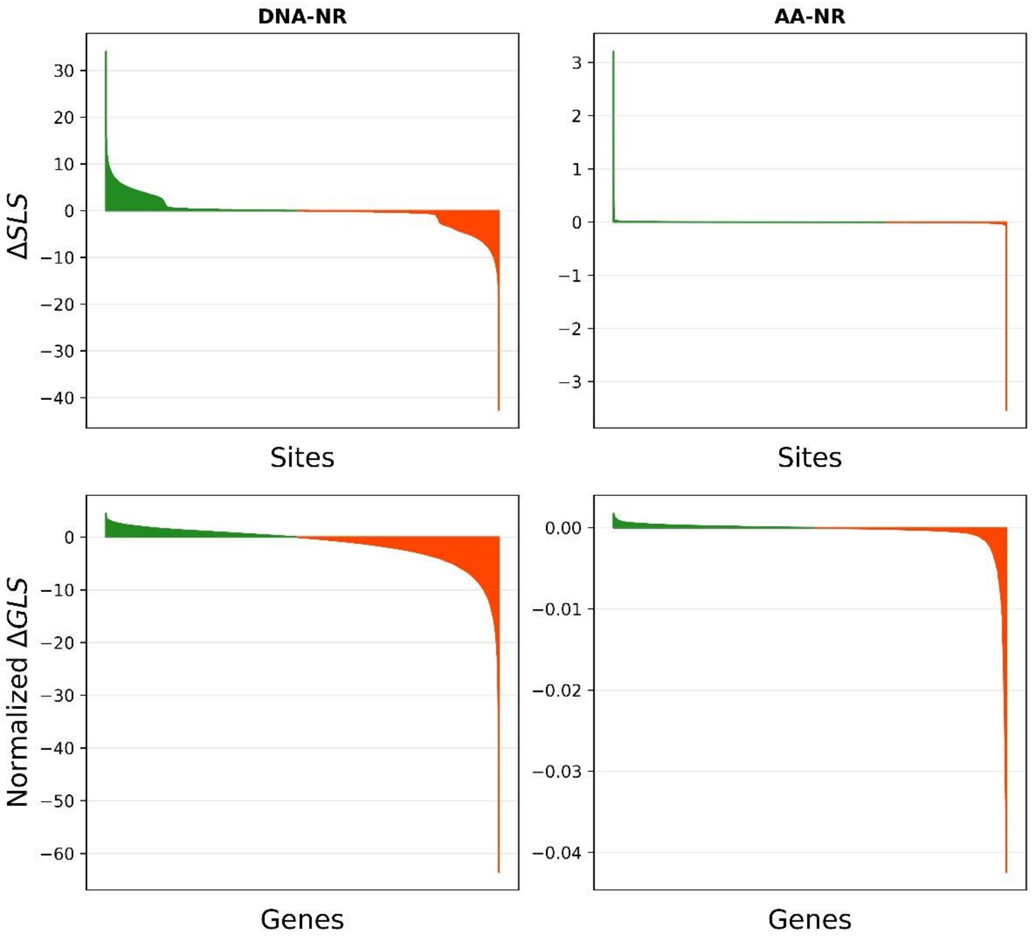


Figure A.5. The normalized difference in the gene-wise log-likelihood score (ΔGLS) and the difference in the site-wise log-likelihood score (ΔSLS) in the Chiroptera amino acid and nucleotide datasets. Positive values (green) present genes/sites that favour the Yinptero-Yango hypothesis and negative values present genes/sites that favour the Micro-Mega hypothesis.

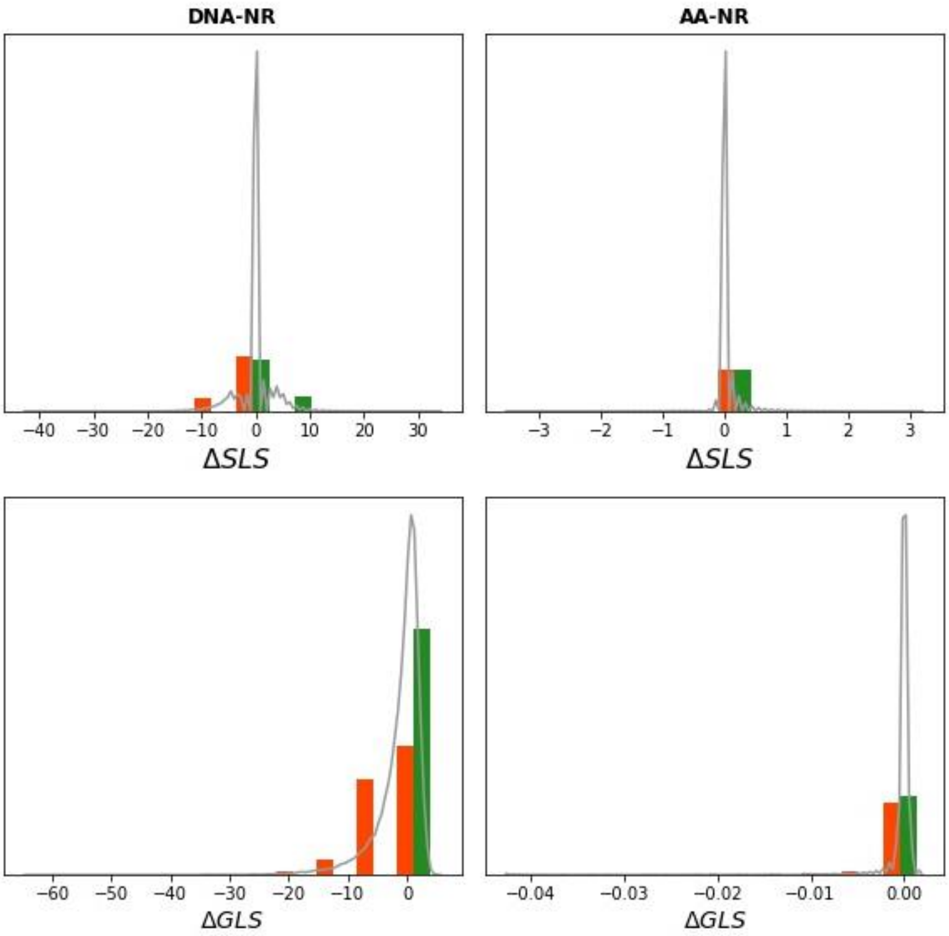


Figure A.6. The distribution of the normalized ΔGLS and ΔSLS in the Chiroptera amino acid and nucleotide datasets where green presents genes that support the Yinptero-Yango hypothesis and orange presents genes that support the Micro-Mega hypothesis.

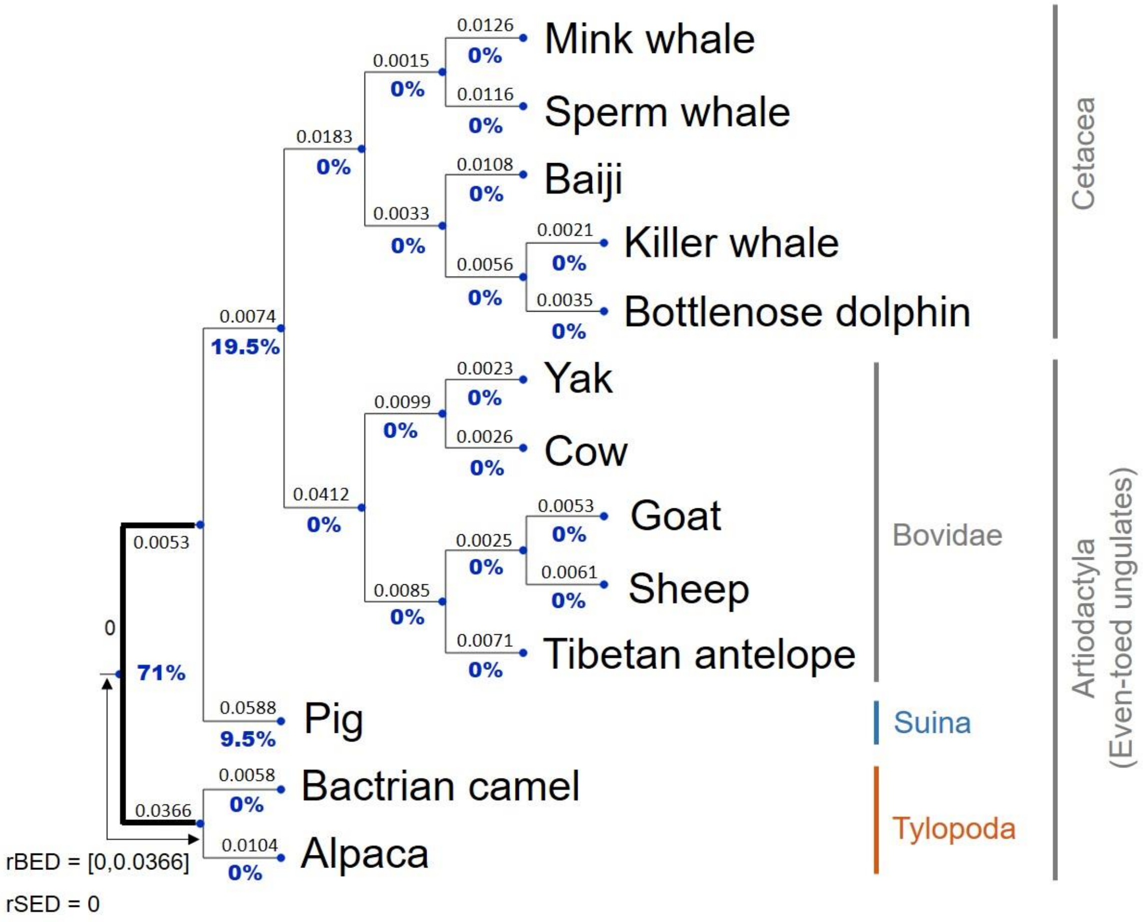


Figure A.7. The ML rooted tree of as inferred from the whole Cetartiodactyla nucleotide dataset. Bold branches are branches in the AU confidence set. Blue values under each branch are the rootstrap support values.

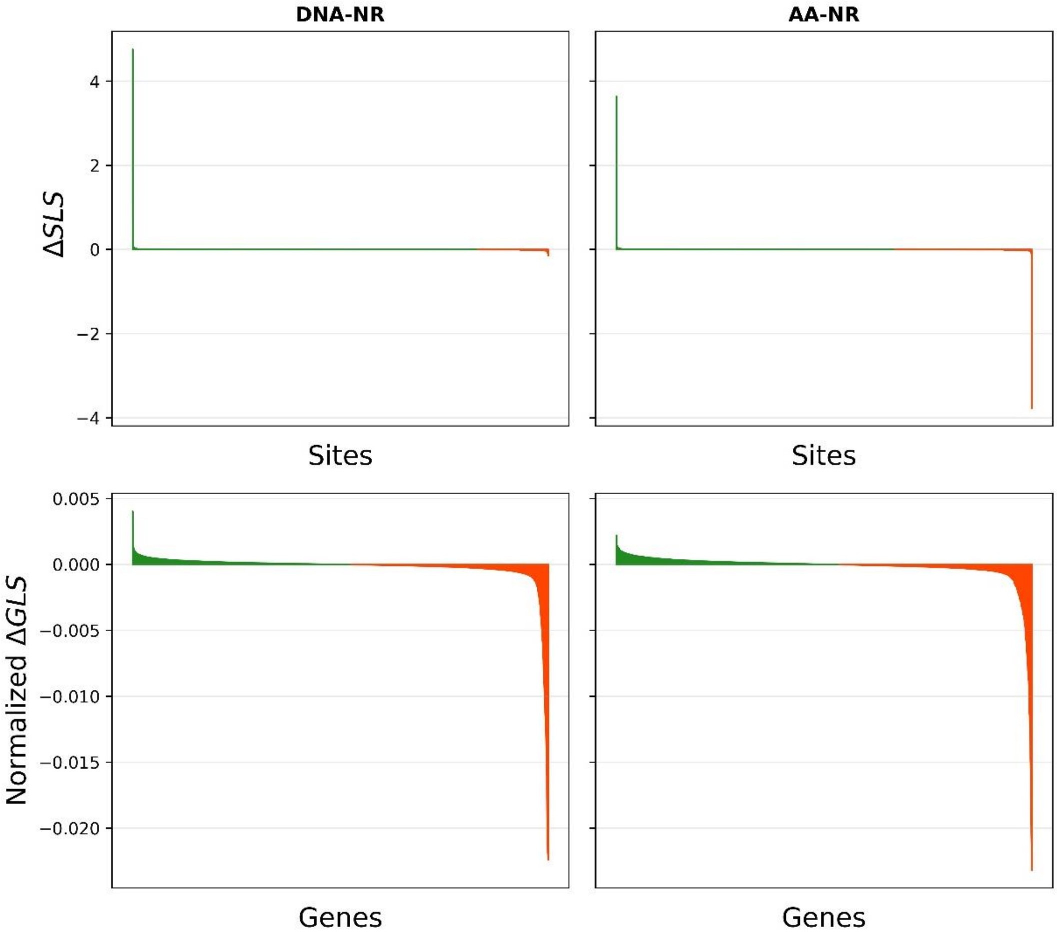


Figure A.8. The normalized difference in the gene-wise log-likelihood score (ΔGLS) and the difference in the site-wise log-likelihood score (ΔSLS) between the Tylopoda+Suina hypothesis and the Tylopoda hypothesis. Positive values (green) present genes/sites that favour the Tylopoda+Suina hypothesis and negative values (orange) present genes/sites that favour the Tylopoda hypothesis.

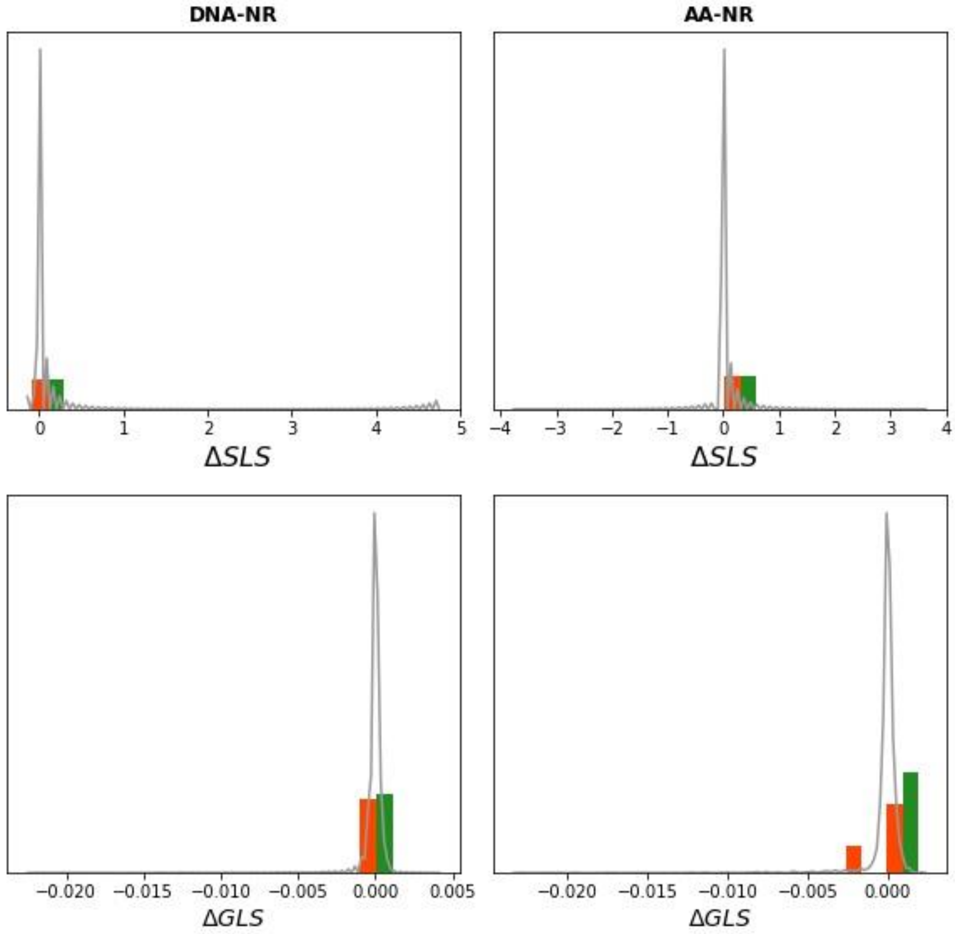


Figure A.9. The distribution of the normalized ΔGLS and ΔSLS in in the Cetartiodactyla amino acid and nucleotide dataset where green presents genes that support the Tylopoda+Suina hypothesis and orange presents genes that support the Tylopoda hypothesis.

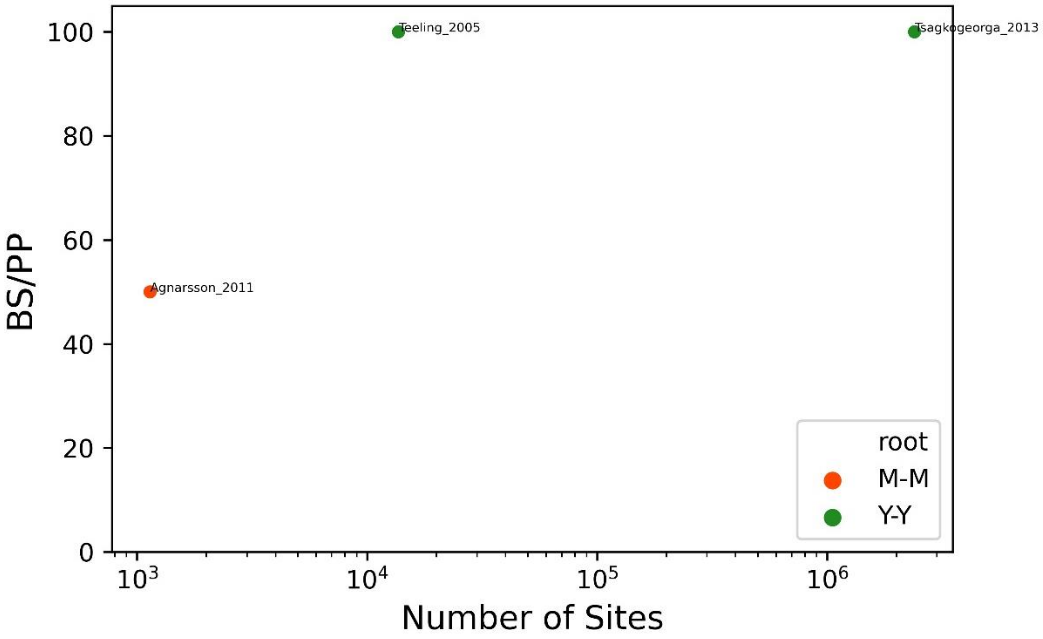


Figure A.10. The support value (bootstrap or posterior probability) for a root placement against the number of nucleotide sites used in each study. Orange dot is for Micro-Mega hypothesis (Agnarsson, et al. 2010) and green dots are for the Yinpterochiroptera-Yangochiroptera hypothesis (Teeling, et al. 2005; Tsagkogeorga, et al. 2013).

References

Agnarsson I, Kuntner M, May-Collado LJ. 2010. Dogs, cats, and kin: a molecular species-level phylogeny of Carnivora. Mol Phylogenet Evol 54:726-745.

Asher RJ, Bennett N, Lehmann T. 2009. The new framework for understanding placental mammal evolution. Bioessays 31:853-864.

Benoit J, Crumpton N, Merigeaud S, Tabuce R. 2013. A memory already like an elephant's? The advanced brain morphology of the last common ancestor of Afrotheria (Mammalia). Brain Behav. Evol. 81:154-169.

Bibi F. 2013. A multi-calibrated mitochondrial phylogeny of extant Bovidae (Artiodactyla, Ruminantia) and the importance of the fossil record to systematics. BMC Evol. Biol. 13:166.

Bibi F, Bukhsianidze M, Gentry AW, Geraads D, Kostopoulos DS, Vrba ES. 2009. The Fossil Record and Evolution of Bovidae: State of the Field. Palaeontologia Electronica 12:1-11.

Blanga-Kanfi S, Miranda H, Penn O, Pupko T, DeBry RW, Huchon D. 2009. Rodent phylogeny revised: analysis of six nuclear genes from all major rodent clades. BMC Evol. Biol. 9:71.

Chatterjee HJ, Ho SY, Barnes I, Groves C. 2009. Estimating the phylogeny and divergence times of primates using a supermatrix approach. BMC Evol. Biol. 9:259.

Chen L, Qiu Q, Jiang Y, Wang K, Lin Z, Li Z, Bibi F, Yang Y, Wang J, Nie W, et al. 2019. Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. Science 364:eaav6202.

Churakov G, Sadasivuni MK, Rosenbloom KR, Huchon D, Brosius J, Schmitz J. 2010. Rodent evolution: back to the root. Mol. Biol. Evol. 27:1315-1326.

Diogo R, Wood B. 2011. Soft-tissue anatomy of the primates: phylogenetic analyses based on the muscles of the head, neck, pectoral region and upper limb, with notes on the evolution of these muscles. J. Anat. 219:273-359.

dos Reis M, Inoue J, Hasegawa M, Asher RJ, Donoghue PC, Yang Z. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. Proc Biol Sci 279:3491-3500.

Eizirik E, Murphy WJ, Koepfli KP, Johnson WE, Dragoo JW, Wayne RK, O'Brien SJ. 2010. Pattern and timing of diversification of the mammalian order Carnivora inferred from multiple nuclear gene sequences. Mol Phylogenet Evol 56:49-63.

Elliot MG, Crespi BJ. 2009. Phylogenetic evidence for early hemochorial placentation in eutheria. Placenta 30:949-967.

Fabre PH, Hautier L, Dimitrov D, Douzery EJ. 2012. A glimpse on the pattern of rodent diversification: a phylogenetic approach. BMC Evol. Biol. 12:88.

Fabre PH, Rodrigues A, Douzery EJ. 2009. Patterns of macroevolution among Primates inferred from a supermatrix of mitochondrial and nuclear DNA. Mol Phylogenet Evol 53:808-825.

Finarelli JA, Flynn JJ. 2009. Brain-size evolution and sociality in Carnivora. Proc Natl Acad Sci U S A 106:9345-9349.

Finstermeier K, Zinner D, Brameier M, Meyer M, Kreuz E, Hofreiter M, Roos C. 2013. A mitogenomic phylogeny of living primates. PLoS One 8:e69504.

Foley NM, Springer MS, Teeling EC. 2016. Mammal madness: is the mammal tree of life not yet resolved? Philos Trans R Soc Lond B Biol Sci 371:20150140.

Gheerbrant E, Amaghzaz M, Bouya B, Goussard F, Letenneur C. 2014. Ocepeia (Middle Paleocene of Morocco): the oldest skull of an afrotherian mammal. PLoS One 9:e89739.

Halliday TJ, Upchurch P, Goswami A. 2015. Resolving the relationships of Paleocene placental mammals. Biol. Rev. Camb. Philos. Soc. 92:521-550.

Hao H, Liu S, Zhang X, Chen W, Song Z, Peng H, Liu Y, Yue B. 2011. Complete mitochondrial genome of a new vole Proedromys liangshanensis (Rodentia: Cricetidae) and phylogenetic analysis with related species: are there implications for the validity of the genus Proedromys? Mitochondrial DNA 22:28-34.

Hartig G, Churakov G, Warren WC, Brosius J, Makalowski W, Schmitz J. 2013. Retrophylogenomics place tarsiers on the evolutionary branch of anthropoids. Sci Rep 3:1756.

Hassanin A, Delsuc F, Ropiquet A, Hammer C, Jansen van Vuuren B, Matthee C, Ruiz-Garcia M, Catzeflis F, Areskoug V, Nguyen TT, et al. 2012. Pattern and timing of diversification of Cetartiodactyla (Mammalia, Laurasiatheria), as revealed by a comprehensive analysis of mitochondrial genomes. C. R. Biol. 335:32-50.

Hedrick BP, Dickson BV, Dumont ER, Pierce SE. 2020. The evolutionary diversity of locomotor innovation in rodents is not linked to proximal limb morphology. Sci Rep 10:717.

Herrera JP, Davalos LM. 2016. Phylogeny and Divergence Times of Lemurs Inferred with Recent and Ancient Fossils in the Tree. Syst. Biol. 65:772-791.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. 35:518-522.

Horn S, Durka W, Wolf R, Ermala A, Stubbe A, Stubbe M, Hofreiter M. 2011. Mitochondrial genomes reveal slow rates of molecular evolution and the timing of speciation in beavers (Castor), one of the largest rodent species. PLoS One 6:e14622.

Jameson NM, Hou ZC, Sterner KN, Weckle A, Goodman M, Steiper ME, Wildman DE. 2011. Genomic data reject the hypothesis of a prosimian primate clade. J. Hum. Evol. 61:295-305.

Kari L, Hill KA, Sayem AS, Karamichalis R, Bryans N, Davis K, Dattani NS. 2015. Mapping the space of genomic signatures. PLoS One 10:e0119815.

Kishino H, Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J. Mol. Evol. 29:170-179.

Kishino H, Miyata T, Hasegawa M. (Kishino1990 co-authors). 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. 31:151-160.

Kumar V, Hallstrom BM, Janke A. 2013. Coalescent-based genome analyses resolve the early branches of the euarchontoglires. PLoS One 8:e60019.

Kuntner M, May-Collado LJ, Agnarsson I. 2011. Phylogeny and conservation priorities of afrotherian mammals (Afrotheria, Mammalia). Zool. Scr. 40:1-15.

Lartillot N, Delsuc F. 2012. Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. Evolution 66:1773-1787.

Liu L, Zhang J, Rheindt FE, Lei F, Qu Y, Wang Y, Zhang Y, Sullivan C, Nie W, Wang J, et al. 2017. Genomic evidence reveals a radiation of placental mammals uninterrupted by the KPg boundary. Proc Natl Acad Sci U S A 114:E7282-E7290.

Machado FA, Zahn TMG, Marroig G. 2018. Evolution of morphological integration in the skull of Carnivora (Mammalia): Changes in Canidae lead to increased evolutionary potential of facial traits. Evolution 72:1399-1419.

Matsui A, Rakotondraparany F, Munechika I, Hasegawa M, Horai S. 2009. Molecular phylogeny and evolution of prosimians based on complete sequences of mitochondrial DNAs. Gene 441:53-66.

Meredith RW, Janecka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simao TL, Stadler T, et al. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. Science 334:521-524.

Minh BQ, Hahn M, Lanfear R. 2018. New methods to calculate concordance factors for phylogenomic datasets. bioRxiv:487801.

Monson TA, Hlusko LJ. 2018. Breaking the rules: Phylogeny, not life history, explains dental eruption sequence in primates. Am. J. Phys. Anthropol. 167:217-233.

Nyakatura K, Bininda-Emonds OR. 2012. Updating the evolutionary history of Carnivora (Mammalia): a new species-level supertree complete with divergence time estimates. BMC Biol. 10:12.

Panciroli E, Janis C, Stockdale M, Martin-Serra A. 2017. Correlates between calcaneal morphology and locomotion in extant and extinct carnivorous mammals. J. Morphol. 278:1333-1353.

Pattinson DJ, Thompson RS, Piotrowski AK, Asher RJ. 2015. Phylogeny, paleontology, and primates: do incomplete fossils bias the tree of life? Syst. Biol. 64:169-186.

Perelman P, Johnson WE, Roos C, Seuanez HN, Horvath JE, Moreira MA, Kessing B, Pontius J, Roelke M, Rumpler Y, et al. 2011. A molecular phylogeny of living primates. PLoS Genet. 7:e1001342.

Phillips MJ, Penny D. 2010. Mammalian Phylogeny. In. Encyclopedia of Life Sciences.

Polly PD, Fuentes-Gonzalez J, Lawing AM, Bormet AK, Dundas RG. 2017. Clade sorting has a greater effect than local adaptation on ecometric patterns in Carnivora. Evol. Ecol. Res. 18:61-95.

Poulakakis N, Stamatakis A. 2010. Recapitulating the evolution of Afrotheria: 57 genes and rare genomic changes (RGCs) consolidate their history. Syst. Biodivers. 8:395-408.

Pozzi L, Hodgson JA, Burrell AS, Sterner KN, Raaum RL, Disotell TR. 2014. Primate phylogenetic relationships and divergence dates inferred from complete mitochondrial genomes. Mol Phylogenet Evol 75:165-183.

Puttick MN, Thomas GH. 2015. Fossils and living taxa agree on patterns of body mass evolution: a case study with Afrotheria. Proc Biol Sci 282:20152023.

Reis MD, Gunnell GF, Barba-Montoya J, Wilkins A, Yang Z, Yoder AD. 2018. Using Phylogenomic Data to Explore the Effects of Relaxed Clocks and Calibration Strategies on Divergence Time Estimation: Primates as a Test Case. Syst. Biol. 67:594-615.

Romiguier J, Ranwez V, Douzery EJ, Galtier N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. Genome Res. 20:1001-1009.

Schenk JJ, Rowe KC, Steppan SJ. 2013. Ecological opportunity and incumbency in the diversification of repeated continental colonizations by muroid rodents. Syst. Biol. 62:837-864.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492-508.

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114-1116.

Springer MS, Meredith RW, Gatesy J, Emerling CA, Park J, Rabosky DL, Stadler T, Steiner C, Ryder OA, Janecka JE, et al. 2012. Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. PLoS One 7:e49521.

Stankowich T, Caro T, Cox M. 2011. Bold coloration and the evolution of aposematism in terrestrial carnivores. Evolution 65:3090-3099.

Steiper ME, Seiffert ER. 2012. Evidence for a convergent slowdown in primate molecular rates and its implications for the timing of early primate evolution. Proc Natl Acad Sci U S A 109:6006-6011.

Svartman M, Stanyon R. 2012. The chromosomes of Afrotheria and their bearing on mammalian genome evolution. Cytogenet. Genome Res. 137:144-153.

Swanson MT, Oliveros CH, Esselstyn JA. 2019. A phylogenomic rodent tree reveals the repeated evolution of masseter architectures. Proc Biol Sci 286:20190672.

Tavares WC, Seuanez HN. 2018. Changes in selection intensity on the mitogenome of subterranean and fossorial rodents respective to aboveground species. Mamm. Genome 29:353-363.

Teeling EC, Springer MS, Madsen O, Bates P, O'Brien S J, Murphy WJ. 2005. A molecular phylogeny for bats illuminates biogeography and the fossil record. Science 307:580-584.

Tomiya S, Tseng ZJ. 2016. Whence the beardogs? Reappraisal of the Middle to Late Eocene 'Miacis' from Texas, USA, and the origin of Amphicyonidae (Mammalia, Carnivora). R Soc Open Sci 3:160518.

Tsagkogeorga G, Parker J, Stupka E, Cotton JA, Rossiter SJ. 2013. Phylogenomic analyses elucidate the evolutionary relationships of bats. Curr. Biol. 23:2262-2267.

Wu J, Hasegawa M, Zhong Y, Yonezawa T. 2014. Importance of synonymous substitutions under dense taxon sampling and appropriate modeling in reconstructing the mitogenomic tree of Eutheria. Genes Genet. Syst. 89:237-251.

Wu J, Yonezawa T, Kishino H. 2017. Rates of Molecular Evolution Suggest Natural History of Life History Traits and a Post-K-Pg Nocturnal Bottleneck of Placentals. Curr. Biol. 27:3025-3033 e3025.

Wu S, Edwards S, Liu L. 2018. Genome-scale DNA sequence data and the evolutionary history of placental mammals. Data Brief 18:1972-1975.

Wu S, Wu W, Zhang F, Ye J, Ni X, Sun J, Edwards SV, Meng J, Organ CL. 2012. Molecular and paleontological evidence for a post-Cretaceous origin of rodents. PLoS One 7:e46445.

Yang C, Xiang C, Qi W, Xia S, Tu F, Zhang X, Moermond T, Yue B. 2013. Phylogenetic analyses and improved resolution of the family Bovidae based on complete mitochondrial genomes. Biochem. Syst. Ecol. 48:136-143.

Yue H, Yan CC, Tu FY, Yang CZ, Ma WQ, Fan ZX, Song ZB, Owens J, Liu SY, Zhang XY. 2015. Two novel mitogenomes of Dipodidae species and phylogeny of Rodentia inferred from the complete mitogenomes. Biochem. Syst. Ecol. 60:123-130.

Zhang ML, Li ML, Ayoola AO, Murphy RW, Wu DD, Shao Y. 2019. Conserved sequences identify the closest living relatives of primates. Zool. Res. 40:532-540.

DISCUSSION

This thesis aims to investigate the prevalence and the effect of model violation due to non-stationary, non-reversible, and non-homogeneous (SRH) evolution on phylogenetic inference and to propose new methods to detect and reduce the bias caused by this model violation. Using mathematical, statistical and bioinformatic tools allowed me to exploit the massive amount of published phylogenetic datasets in order to explore the extent of model violation in empirical data and to develop new tests to address this issue.

Based on Bowker (Bowker 1948), Stuart (Stuart 1955), and Ababneh's (Ababneh, et al. 2006) test statistics; i.e. the matched-pairs tests of homogeneity (Jermiin, et al. 2017), I developed the MaxSym tests to accommodate multiple-sequence alignments to check for model violation due to non-SRH evolution in a representative sample of 35 published datasets with more than 3500 genome partitions. I found that model violation is more widespread than had been previously believed and it has an enormous impact on the phylogenetic inference. The model violation is mainly prevalent in the third codon positions of most types of genomes (nuclear, mitochondrial, and viral) and intergenic spacers in plastids. Yet, the results confirm that all types of genomes and genomic regions are prone to SRH violation.

A number of studies have used these new tests since their publication. For example, using the MaxSym tests to remove loci that have a high potential to introduce systematic bias and compromise the consistency of the phylogenetic inference facilitated the resolution of problematic phylogenies such as Terebelliformia (Stiller, et al. 2020), Hyaenidae (Westbury, et al. 2021), bioluminescent Elateroids (Kusy, et al. 2021), Gracillariidae (Li, et al. 2021), Pteropodidae (Nesi, et al. 2021), Fungiidae (Grinblat, et al. 2021) and provided new insights to the current knowledge regarding their evolution.

A deeper look into the results of the three MaxSym tests from the empirical dataset in Chapter 1 reveals that most of the model violations in these datasets are due to non-stationary evolution. Although non-homogeneous evolution is common in phylogenetics, apparently it is not the main source of systematic bias due to model violation in phylogenetic inference (Jayaswal, et al. 2005; Ababneh, et al. 2006; Song, et al. 2010). These results suggest that developing non-stationary substitution models that are efficient and user-friendly could significantly improve phylogenetic inference by reducing systematic bias due to non-SRH evolution.

To investigate further the impact of systematic bias due to the use of SRH models on data that has not evolved under such conditions, I simulated data under various settings and different sizes. In the convergent scheme simulations where two distantly related branches undergo correlated and extreme changes in the substitution models, the phylogenetic inference was rigorously biased when SRH models were used for the phylogenetic analysis. However, under the inheritance scheme where branches inherit their substitution models from their ancestors, the phylogenetic inference was robust to model violations.

Although it is tempting to conclude from these results that phylogenetic inference is robust to model violation when these violations are not severe, such a conclusion should be drawn with lots of caution. Even though I attempted to simulate data under realistic conditions, those simulations still have many limitations that make them far from realistic. To name some, the empirical distributions of base frequencies, substitution rates and branch lengths were all derived using SRH models, specifically, GTR models, that constrain the estimation of the evolutionary process of the empirical datasets. While these empirical distributions are more realistic than distributions that are commonly used in simulations in phylogenetics, they oversimplify the real evolutionary processes.

Another limitation is that the branch lengths distribution was derived from trees that were inferred using SRH models and without distinguishing between internal and leaf branches. Moreover, by definition branches reach their stationary distribution when time approaches infinity. However, in my simulated data, I used branch lengths drawn from the empirical distribution of branch lengths. In other words, as long as I do not have infinite branch lengths, the simulated alignments could be much less non-stationary than anticipated.

Despite not being perfectly realistic, the empirical distributions derived from real datasets, along with their best-fit probability distributions can be are very useful for numerous purposes, such as simulations or prior distributions for Bayesian analysis. These distributions are available in the new simulator, AliSim (Ly-Trong, et al. 2021), and can be used to generate biologically realistic alignments.

Testing the power of the MaxSym tests using these simulations reveals that their power is rather limited. This is not surprising since these tests consider only the most divergent pair of sequences while ignoring all the other sequences in the alignment. Regardless, the false-positive rates of those tests are somewhat reasonable which makes them suitable as an elimination method for loci that violates the SRH assumption. One alternative for the MaxSym tests is using all the possible pairs of sequences instead of just the pair with the maximum divergence. This will significantly increase the power of the test to detect partitions that violate the SRH assumptions. The only downside with this approach is the dependencies between the pairs of sequences, which makes it statistically improper.

In an empirical framework, the dependencies between the pairs of sequences have little effect on the desired result. Accounting for these dependencies may increase type I error due to over-estimation of model violation, which could lead to rejecting parts of the data that do not violate the SRH assumption. However, in the era of big data where the size of datasets are on a genome-scale, using only a little part of the data for the phylogenetic analysis is completely acceptable and sometimes even desired (Philippe, et al. 2011; Kumar, et al. 2012; Yang and Zhu 2018). In fact, methods that use partial datasets, such as cross-validation, is currently very common in phylogenetic studies (Susko and Roger 2020) although they can also suffer from dependency issues (Wong and Yang 2017).

One suggestion as a new extension for the three matched-pairs tests of symmetry (Jermiin, et al. 2017) to accommodate multiple sequence alignments with high power would be a binomial test for all the pairwise p-values in the alignment. Applying the matched-pair tests of symmetry to every pair of the sequences would result in(n¦2) chi-squared p-values, where n is the number of sequences in the alignment. Then using a binomial distribution with (n¦2) trials and 5% success probability we can assess if the alignment passes or rejects the null hypothesis of SRH evolution. Indeed, this is the approach that I first proposed when working on this problem but it was rejected by reviewers during the review process as it is impossible to account for the non-independence between the pairs of sequences without a priori knowledge of the phylogeny. While this approach ignores the dependencies among the pairwise p-values, it is much more powerful than the MaxSym test in detecting model violations due to non-SRH evolution in multiple sequence alignments.

Assessing the model assumptions a priori to the phylogenetic inference is a very important step and therefore it is part of the phylogenetic protocol proposed by (Jermiin, et al. 2020). In this thesis, I addressed this step by introducing the MaxSym tests that can be used on multiple sequence alignments to exclude partitions that violates the SRH assumptions. Another important step that was suggested by (Jermiin, et al. 2020) is using tests for goodness-of-fit a posterior to the phylogenetic inference. To address this step, I developed a new algorithm that checks the optimal number of substitution models operating along the phylogenetic tree given pre-defined clades for that phylogeny.

The results of using this new algorithm on three big datasets indicate that the homogeneous model with one base stationary distribution and one substitution matrix operating along the whole tree always has the worst fit for the data. These results are congruent with results from previous studies that compared homogeneous and non-homogeneous models such as (Herbeck, et al. 2005; Blanquart and Lartillot 2006; Boussau and Gouy 2006; Blanquart and Lartillot 2008; Boussau, et al. 2008; Dutheil and Boussau 2008; Jayaswal, et al. 2011; Zhang, et al. 2011; Dutheil, et al. 2012; Groussin, et al. 2013; Jayaswal, et al. 2014) and showed that non-homogeneous models always outperform homogeneous models according to statistical criteria such as AIC and BIC.

The major limitation of using non-homogeneous models in phylogenetic studies is efficiency, especially in amino acid datasets. The large number of parameters and the risk of over-parameterization makes the use of these models impractical in empirical datasets. By utilizing the efficiency of QMaker (Minh, et al. 2021) to estimate amino acid substitution models and using pre-defined homogeneous clades, I introduced a user-friendly algorithm for non-homogeneous model inference. Although this approach is far from perfect, it provided statistically improved models of evolution than the commonly used SRH models with datasets that are tens of folds larger than have been previously used with non-homogeneous amino acid models (Groussin, et al. 2013).

Using a non-homogeneous and non-stationary model showed that the differences between stationary frequencies are more pronounced than differences between Q matrices within major clades, which confirms what I found using the MaxSym tests on various empirical datasets in Chapter 1. Although varying the stationary distribution across the tree could be enough to significantly improve the model fitness without varying the substitution processes, in my approach I did not have this option. Though, in another study by (Groussin, et al. 2013), they tried three different types of models – homogeneous and stationary, homogeneous and non-stationary, and non-homogeneous and non-stationary on four empirical datasets and found that in one of the datasets the homogeneous and non-stationary model was the best-fit model while in the remaining three datasets the non-homogeneous and non-stationary model was the best-fit model in terms of BIC score. These results suggest that while in some cases the non-stationary model is indeed the best-fit model, in most cases the non-homogeneous and non-stationary model will outperform it.

In addition to the stationarity and homogeneity assumptions that I extensively explored in the first three chapters, the third assumption of SRH evolution, namely; reversibility, was not addressed. Hence, in the fourth chapter of this thesis, I focused on investigating non-reversible models of nucleotide and amino acid substitutions and comparing them to the commonly used time-reversible models. In terms of BIC score, the results demonstrate a clear preference for the non-reversible models of amino acids as the number of sites increases in the dataset. Yet, in the non-reversible models of nucleotides, this trend was not there. On the contrary, in most datasets, the reversible model had the best fit for the data regardless of the number of sites. This is likely due to the fact that I used partitioned models in the reversible analysis but not in the non-reversible analysis.

Since the “Pulley Principle” (Felsenstein 1981) that allows moving the root of the tree without affecting its likelihood does not hold for non-reversible models, inferring rooted trees without additional assumptions or information is a powerful feature of those models. The results in Chapter 4 present the ability of the non-reversible models of nucleotides and amino acids to correctly infer the root placement of most trees. Furthermore, in order to assess the confidence in the root placement, I developed the rootstrap support value which is a statistical measure of the support that the data have for the placement of the root in rooted trees. Applying that measure for the inferred rooted trees shows very high support for the correct root placement in whole datasets. As expected, due to a large number of parameters in the non-reversible models, the rootstrap support value decreases with decreasing the number of sites in the alignment.

In Chapter 3 I found that the two clades; Chiroptera and Cetartiodactyla are homogeneous, meaning that the subclades in each clade have similar evolutionary processes. Yet, an interesting fact about those two clades is that their root placements are still controversial. Therefore, I used the non-reversible models of nucleotides and amino acids to shed more light on that problem. The results show that for the Chiroptera clade, both the nucleotide and the amino acid models place the root on the branch connecting between Yinpterochiroptera and Yangochiroptera and with very high rootstrap support in the nucleotide dataset. This result reinforces the Yango-Yinptero hypothesis that is gaining more and more support in recent studies (e.g. Meganathan, et al. 2012; Tsagkogeorga, et al. 2013; Ren, et al. 2018; Reyes-Amaya and Flores 2019). For Cetartiodactyla, the nucleotide's model inferred Tylopoda as the most-basal clade while the amino acid model inferred the clade that contains Tylopoda and Suina as the most basal clade of Cetartiodactyla. In both cases, the rootstrap support was not very high (~71%) but very similar.

The use of non-reversible models is proved to be very useful, especially when no additional information or assumptions are available or desired for rooting the tree. The computational load of these models is reasonable even with very large datasets like the ones I used in this study. The amino acid non-reversible models, in particular, seem to always outperform the traditional time-reversible models in terms of BIC scores when the number of sites is large enough, suggesting that the use of these models is recommended for large datasets, even when rooting the inferred tree is not a primary goal of the analysis.

The work presented in this thesis highlights the complex nature of phylogenetic inference and emphasize the need for appropriate methods and tools to accommodate this complexity. Relying on traditional stationary, reversible, and homogeneous substitution models to explore evolutionary relationships in the era of genomic-scale datasets might be convenient but it is very misleading. The literature is crammed with examples of controversies caused by the inconsistency of phylogenetic inference due to bias introduced by model violations. The mutational process is known to occur heterogeneously along genomes. For example, the rate of C to T in mammalian genomes is much higher for CpG dinucleotides than for other dinucleotides due to the higher DA methylation level in these sites (Ehrlich and Wang 1981). Another example is the antiviral activities of the members of APOBEC cytidine deaminases. Those proteins induce the hypermutation G to A in the HIV genome, as well as the hypermutation C to U in the SARS-CoV-2 genome (Simmonds 2020; Wang, et al. 2020) and the hypermutations G to A and C to T in the latest monkeypox virus outbreak (Gomes, et al. 2022).

Since “essentially, all models are wrong, but some are useful” (Box 1979), verifying that the model assumptions are not violated by the data is an important step forward in minimizing bias in the phylogenetic analysis. I believe that there is still more to be done in that regard.

Another avenue that is worth investing in, is the development of more complex models that on one hand can contain the complexity of the real data and on the other hand be efficient and user-friendly to gain popularity among researchers. The non-reversible models that I investigated in this thesis demonstrate the ability of such models to improve phylogenetic inference without extra hurdles for the user. Another one that I also explored, is the non-stationary and non-homogeneous model. An inclusive model that allows for non-stationary, non-reversible, and non-homogeneous substitution processes should be the logical next step in phylogenetic inference.

In this thesis, I focused on the SRH assumptions, their prevalence in biological datasets, the impact that their violation has on the phylogenetic inference and introduced new methods to deal with the implications. However, some of the most fundamental assumptions in phylogenetics were not addressed in this work. One of the basic assumptions that almost all phylogenetic analysis starts with is that the substitution processes operating along phylogenies are Markovian. It is so natural to assume the Markovian property for evolutionary processes, that a very small number of studies even tried to investigate the violation of this assumption (Tuller and Mossel 2010; Vera-Ruiz, et al. 2014; Vera-Ruiz, et al. 2021).

Another assumption that seems very natural, is the assumption that the sites of the alignments evolve independently from each other and with identical distribution (i.i.d.). Despite the growing body of evidence that substitution processes are context-dependent and affected by neighbouring sites, substitution models that account for dependencies between sites are still rarely used in phylogenetic analysis (Pedersen and Jensen 2001; Arndt, et al. 2003; Lunter and Hein 2004; Siepel and Haussler 2004; Arndt and Hwa 2005; Christensen 2006; Shapiro, et al. 2006; Baele, et al. 2008; Bérard, et al. 2008; Hobolth 2008; Baele, et al. 2010; Nasrallah, et al. 2011; Bérard and Guéguen 2012).

Following the phylogenetic protocol proposed by (Jermiin, et al. 2020) is an important step towards reducing systematic bias in phylogenetic inference. In my thesis I tried to address the two new steps suggested in this protocol, namely, assessing phylogenetic assumptions a priori to the phylogenetic inference and testing for goodness-of-fit a posteriori to the phylogenetic inference. Moreover, I demonstrated the advantages of using more complex models of evolution to relax some of the common assumptions in phylogenetics. While this protocol is a good starting point for a more robust phylogenetic inference, more work is needed to develop better methods that can keep in line with the rapid advancement in DNA sequencing and data collecting.

References

Ababneh F, Jermiin LS, Ma C, Robinson J. 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics 22:1225-1231.

Arndt PF, Burge CB, Hwa T. 2003. DNA sequence evolution with neighbor-dependent mutation. J. Comput. Biol. 10:313-322.

Arndt PF, Hwa T. 2005. Identification and measurement of neighbor-dependent nucleotide substitution processes. Bioinformatics 21:2322-2328.

Baele G, Van de Peer Y, Vansteelandt S. 2008. A model-based approach to study nearest-neighbor influences reveals complex substitution patterns in non-coding sequences. Syst. Biol. 57:675-692.

Baele G, Van de Peer Y, Vansteelandt S. 2010. Using Non-Reversible Context-Dependent Evolutionary Models to Study Substitution Patterns in Primate Non-Coding Sequences. J. Mol. Evol. 71:34-50.

Bérard J, Gouéré J-B, Piau D. 2008. Solvable models of neighbor-dependent substitution processes. Math. Biosci. 211:56-88.

Bérard J, Guéguen L. 2012. Accurate Estimation of Substitution Rates with Neighbor-Dependent Models in a Phylogenetic Context. Syst. Biol. 61:510-521.

Blanquart S, Lartillot N. 2006. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Mol. Biol. Evol. 23:2058-2071.

Blanquart S, Lartillot N. 2008. A site- and time-heterogeneous model of amino acid replacement. Mol. Biol. Evol. 25:842-858.

Boussau B, Blanquart S, Necsulea A, Lartillot N, Gouy M. 2008. Parallel adaptations to high temperatures in the Archaean eon. Nature 456:942-945.

Boussau B, Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. Syst. Biol. 55:756-768.

Bowker AH. 1948. A test for symmetry in contingency tables. J Am Stat Assoc 43:572-574.

Box GEP. 1979. Robustness in the Strategy of Scientific Model Building. In. Robustness in Statistics. p. 201-236.

Christensen F. 2006. Pseudo-likelihood for Non-reversible Nucleotide Substitution Models with Neighbour Dependent Rates. In. Statistical Applications in Genetics and Molecular Biology.

Dutheil J, Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. BMC Evol. Biol. 8:255.

Dutheil JY, Galtier N, Romiguier J, Douzery EJ, Ranwez V, Boussau B. 2012. Efficient selection of branch-specific models of sequence evolution. Mol. Biol. Evol. 29:1861-1874.

Ehrlich M, Wang RY. 1981. 5-Methylcytosine in eukaryotic DNA. Science 212:1350-1357.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368-376.

Gomes JP, Isidro J, Borges V, Pinto M, Sobral D, Santos J, Mixão V, Ferreira R, Nunes A, Santos D. 2022. Multi-country outbreak of monkeypox virus: phylogenomic characterization and signs of microevolution.

Grinblat M, Cooke I, Shlesinger T, Ben-Zvi O, Loya Y, Miller DJ, Cowman PF. 2021. Biogeography, reproductive biology and phylogenetic divergence within the Fungiidae (mushroom corals). Mol. Phylogen. Evol. 164:107265.

Groussin M, Boussau B, Gouy M. 2013. A branch-heterogeneous model of protein evolution for efficient inference of ancestral sequences. Syst. Biol. 62:523-538.

Herbeck JT, Degnan PH, Wernegreen JJ. 2005. Nonhomogeneous model of sequence evolution indicates independent origins of primary endosymbionts within the enterobacteriales (gamma-Proteobacteria). Mol. Biol. Evol. 22:520-532.

Hobolth A. 2008. A Markov chain Monte Carlo expectation maximization algorithm for statistical analysis of DNA sequence evolution with neighbor-dependent substitution rates. Journal of Computational and Graphical Statistics 17:138-162.

Jayaswal V, Jermiin LS, Poladian L, Robinson J. 2011. Two stationary nonhomogeneous Markov models of nucleotide sequence evolution. Syst. Biol. 60:74-86.

Jayaswal V, Jermiin LS, Robinson J. 2005. Estimation of Phylogeny Using a General Markov Model. Evol Bioinform 1:62-80.

Jayaswal V, Wong TK, Robinson J, Poladian L, Jermiin LS. 2014. Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. Syst. Biol. 63:726-742.

Jermiin LS, Catullo RA, Holland BR. 2020. A new phylogenetic protocol: dealing with model misspecification and confirmation bias in molecular phylogenetics. NAR Genom Bioinform 2:lqaa041.

Jermiin LS, Jayaswal V, Ababneh FM, Robinson J. 2017. Identifying Optimal Models of Evolution. In: Keith JM, editor. Bioinformatics. Melbourne: Humana Press, New York, NY. p. 379-420.

Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K. 2012. Statistics and truth in phylogenomics. Mol. Biol. Evol. 29:457-472.

Kusy D, He JW, Bybee SM, Motyka M, Bi WX, Podsiadlowski L, Li XY, Bocak L. 2021. Phylogenomic relationships of bioluminescent elateroids define the ‘lampyroid’clade with clicking Sinopyrophoridae as its earliest member. Syst. Entomol. 46:111-123.

Li X, St Laurent R, Earl C, Doorenweerd C, van Nieukerken EJ, Davis DR, Johns CA, Kawakita A, Kobayashi S, Zwick A, et al. 2021. Phylogeny of gracillariid leaf-mining moths: evolution of larval behaviour inferred from phylogenomic and Sanger data. Cladistics n/a.

Lunter G, Hein J. 2004. A nucleotide substitution model with nearest-neighbour interactions. Bioinformatics 20:i216-i223.

Ly-Trong N, Naser-Khdour S, Lanfear R, Minh BQ. 2021. AliSim: A Fast and Versatile Phylogenetic Sequence Simulator For the Genomic Era. bioRxiv:2021.2012.2016.472905.

Meganathan PR, Pagan HJ, McCulloch ES, Stevens RD, Ray DA. 2012. Complete mitochondrial genome sequences of three bats species and whole genome mitochondrial analyses reveal patterns of codon bias and lend support to a basal split in Chiroptera. Gene 492:121-129.

Minh BQ, Dang CC, Vinh LS, Lanfear R. 2021. QMaker: Fast and Accurate Method to Estimate Empirical Models of Protein Evolution. Syst. Biol. 70:1046-1060.

Nasrallah CA, Mathews DH, Huelsenbeck JP. 2011. Quantifying the impact of dependent evolution among sites in phylogenetic inference. Syst. Biol. 60:60-73.

Nesi N, Tsagkogeorga G, Tsang SM, Nicolas V, Lalis A, Scanlon AT, Riesle-Sbarbaro SA, Wiantoro S, Hitch AT, Juste J, et al. 2021. Interrogating Phylogenetic Discordance Resolves Deep Splits in the Rapid Radiation of Old World Fruit Bats (Chiroptera: Pteropodidae). Syst. Biol.

Pedersen A-MK, Jensen JL. 2001. A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. Mol. Biol. Evol. 18:763-776.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DT, Manuel M, Worheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.

Ren M, Sun HJ, Bo SQ, Zhang SY, Hua PY. 2018. Parallel amino acid deletions of prestin protein in two dramatically divergent bat lineages suggest the complexity of the evolution of echolocation in bats. Acta Chiropterologica 20:311-317.

Reyes-Amaya N, Flores D. 2019. Hypophysis size evolution in Chiroptera. Acta Chiropterologica 21:65-74.

Shapiro B, Rambaut A, Drummond AJ. 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol. Biol. Evol. 23:7-9.

Siepel A, Haussler D. 2004. Phylogenetic Estimation of Context-Dependent Substitution Rates by Maximum Likelihood. Mol. Biol. Evol. 21:468-488.

Simmonds P. 2020. Rampant C-->U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. mSphere 5:2020.2005.2001.072330.

Song H, Sheffield NC, Cameron SL, Miller KB, Whiting MF. 2010. When phylogenetic assumptions are violated: base compositional heterogeneity and among-site rate variation in beetle mitochondrial phylogenomics. Syst. Entomol. 35:429-448.

Stiller J, Tilic E, Rousset V, Pleijel F, Rouse GW. 2020. Spaghetti to a Tree: A Robust Phylogeny for Terebelliformia (Annelida) Based on Transcriptomes, Molecular and Morphological Data. Biology (Basel) 9:73.

Stuart A. 1955. A Test for Homogeneity of the Marginal Distributions in a Two-Way Classification. Biometrika 42:412-416.

Susko E, Roger AJ. 2020. On the use of information criteria for model selection in phylogenetics. Mol. Biol. Evol. 37:549-562.

Tsagkogeorga G, Parker J, Stupka E, Cotton JA, Rossiter SJ. 2013. Phylogenomic analyses elucidate the evolutionary relationships of bats. Curr. Biol. 23:2262-2267.

Tuller T, Mossel E. 2010. Co-evolution is incompatible with the markov assumption in phylogenetics. IEEE/ACM transactions on computational biology and bioinformatics 8:1667-1670.

Vera-Ruiz VA, Lau KW, Robinson J, Jermiin LS editors. BMC Bioinformatics. 2014.

Vera-Ruiz VA, Robinson J, Jermiin LS. 2021. A Likelihood-Ratio Test for Lumpability of Phylogenetic Data: Is the Markovian Property of an Evolutionary Process retained in Recoded DNA? Syst. Biol.

Wang R, Hozumi Y, Zheng Y-H, Yin C, Wei G-W. 2020. Host Immune Response Driving SARS-CoV-2 Evolution. Viruses 12:1095.

Westbury MV, Le Duc D, Duchene DA, Krishnan A, Prost S, Rutschmann S, Grau JH, Dalen L, Weyrich A, Noren K, et al. 2021. Ecological Specialization and Evolutionary Reticulation in Extant Hyaenidae. Mol. Biol. Evol. 38:3884-3897.

Wong T-T, Yang N-Y. 2017. Dependency analysis of accuracy estimates in k-fold cross validation. IEEE Transactions on Knowledge and Data Engineering 29:2417-2427.

Yang Z, Zhu T. 2018. Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees. Proc Natl Acad Sci U S A 115:1854-1859.

Zhang C, Wang J, Xie W, Zhou G, Long M, Zhang Q. 2011. Dynamic programming procedure for searching optimal models to estimate substitution rates based on the maximum-likelihood method. Proc Natl Acad Sci U S A 108:7860-7865.