The Influence of Model Violation on Phylogenetic Inference: A Simulation Study

SUHA NASER-KHDOUR*[1], BUI QUANG MINH[2], AND ROBERT LANFEAR*[1]

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

[2] School of Computing, Australian National University, Canberra, Australian Capital Territory, Australia

**Authors for Correspondence:* rob.lanfear@anu.edu.au; suha.naser@alumni.anu.edu.au

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]

MAIN TEXT

Markov models are commonly used in model-based phylogenetic analyses such as maximum likelihood (ML) and Bayesian inference (Felsenstein 2004; Yang 2006). A Markov model is often 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 assumption of evolution under Markovian conditions 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). Such approaches can be made more mathematically and computationally tractable by assuming that the Markov model is stationary, reversible, and homogeneous (SRH), assumptions which are universal to many of the most commonly-used models in phylogenetics such as those in the GTR family (Kimura 1980; Felsenstein 1981; Hasegawa, et al. 1985; Tavaré 1986; Tamura and Nei 1993; Yang 1994). Time-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 substitution process remains the same in both directions.

The assumptions of homogeneity, stationarity, and reversibility imply a potential departure from biological reality (Yang and Roberts 1995; Foster and Hickey 1999; Foster 2004; Ababneh, et al. 2006). Various evolutionary processes, including mutational biases, compositional heterogeneity driven by factors such as GC-biased gene conversion, and changes in selective pressures leading to amino acid compositional shifts, on phylogenetic inference under the assumption of stationarity, reversibility, and homogeneity. Hypermutations, such as the CpG islands in mammals (Bird 1980) is an example of a non-reversible model of evolution. 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 (Yang 1994). These degrees of freedom can increase dramatically if the model is non-stationary or/and non-homogeneous (Jayaswal, et al. 2005; Jayaswal, et al. 2011a; Jayaswal, et al. 2011b; Jayaswal, et al. 2014) 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 for large datasets. On the other hand, removing data that do not comply with the SRH assumption, as suggested by Jermiin et al. (2017), will come at the cost of losing phylogenetic information. Both studies of simulated data (Huelsenbeck and Hillis 1993; Hillis, et al. 1994; Galtier and Gouy 1998; Conant and Lewis 2001; Ho and Jermiin 2004; Jermiin, et al. 2004; Boussau and Gouy 2006) and empirical data (Galtier and Gouy 1995; Galtier and Gouy 1998; Phillips, et al. 2004; Collins, et al. 2005; Jayaswal, et al. 2005; Jayaswal, et al. 2007; Phillips, et al. 2010; Jayaswal, et al. 2011b; Nguyen, et al. 2012; Betancur-R, et al. 2013; Naser-Khdour, et al. 2019) studies have shown that applying SRH models to data that have evolved under more complex conditions may lead to significant errors in phylogenetic inference. However, most of these simulation studies have used

parameters or parameter values that are not likely to 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 (e.g. Hillis, et al. 1994; Ho and Jermiin 2004; Jermiin, et al. 2004). 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 (e.g. Jermiin et al. 2004) 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 and parameter values 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. It is noteworthy that by non-homogeneous conditions we refer to heterogeneity across lineages and not heterogeneity across sites. In contrast to heterogeneity across lineages that are widely ignored in phylogenetic analyses, heterogeneity across sites is usually accounted for by using a pre-defined distribution of rate variation amongst sites, such as Gamma distribution (Yang 1994). For the purpose of this study, we assume a constant rate of substitution among sites.

MATERIALS AND METHODS

Code and Data Availability

All scripts and code used to perform the analyses in this study are available on GitHub at: https://github.com/suhanaser/empiricalGTRdist.

Simulated sequence alignments and associated metadata are deposited in the Dryad Digital Repository, DOI: 10.5061/dryad.k3j9kd582.

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 $\Phi = \{1, \ldots, l\}$.

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. The code used to perform these simulations is available in the inheritance_Simulations.py script of our GitHub repository.
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 that is distinct from the substitution process that operates on the rest of the tree. This scheme resembles convergent evolution. The code used to perform these simulations is available in the two_matrix_Simulations.py script of our GitHub repository.

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 loci (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 (Minh, et al. 2020) with a GTR model and free rate heterogeneity across sites (Yang 1995) with 4 categories + invariant sites and extracted the parameters using the script extract_GTR_parameters_dist_from_empirical_data.py script available on our GitHub repository. 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). Assuming that sampling among each distribution independently will approximate a sample from the joint distribution of parameters, 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) using the run birth-death.R available on out GitHub repository. 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 ω and v 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 the state frequency to vary 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, π_0 is the stationary base frequencies, π_{root} is the base frequencies at the root, and d_{root} is the branch length of the root branch.

When $\nu = 1$, π_{root} is equal to π_0 and this scheme boils down to the first SRH condition. When $\nu = 0$, 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:

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

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:

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

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 (π_{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.

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

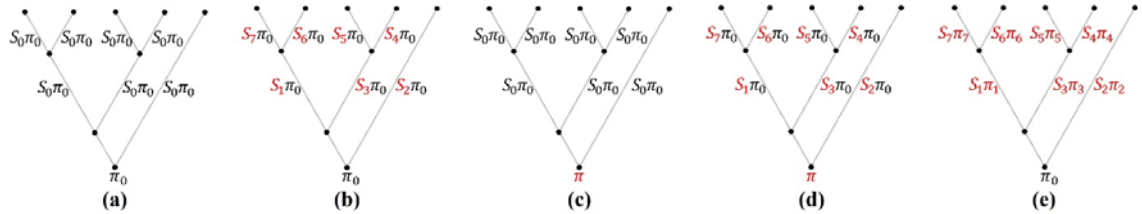

FIGURE 1. An example of 5 taxon tree with different degrees for homogeneity and stationarity. **(a)** stationary and homogeneous, **(b)** stationarity but not homogeneous, **(c)** non-stationary but homogeneous, **(d)** non-stationary and non-homogeneous where the stationarity and homogeneity assumptions 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. 2FIGURE 2).

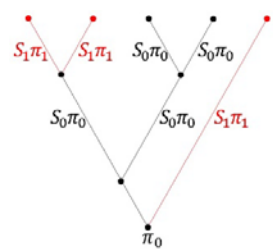

FIGURE 2. Non-stationary and non-homogeneous process with two different rate matrices Q_0 and Q_1.

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 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 splits 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.

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 consisting 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 1, Appendix Table A.2, Figs. A.1-3).

TABLE 1. 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	-5.4e-31
Branch length	Power-log-normal	1.208, 1.443	0.017	-7.1 e-05
Internal	Log-normal	1.272	0.010	-1.5 e-04
Leaves	Power-log-normal	2.805, 1.838	0.076	3.1 e-05

Furthermore, we checked if the distributions of the internal branch lengths and the leaf branch lengths are similar. Our results show the distribution of the leaf branch lengths is similar to the general distribution of all branches but with different parameters. On the other hand, the distribution of the internal branch lengths is different (Appendix Fig. A.3). Yet, for the purpose of this study, we will use the general branch length distribution for all branches.

Both Simulation Schemes Capture the True extent of Compositional Heterogeneity

In order to confirm that our simulation settings are producing compositional variance patterns that are of the same order of magnitude as what we observed in the empirical data, we compared the variance in GC content between the simulated datasets and the empirical datasets.

Our results confirm that the compositional variance in both simulation schemes is indeed similar to the empirical data (Fig. 3).

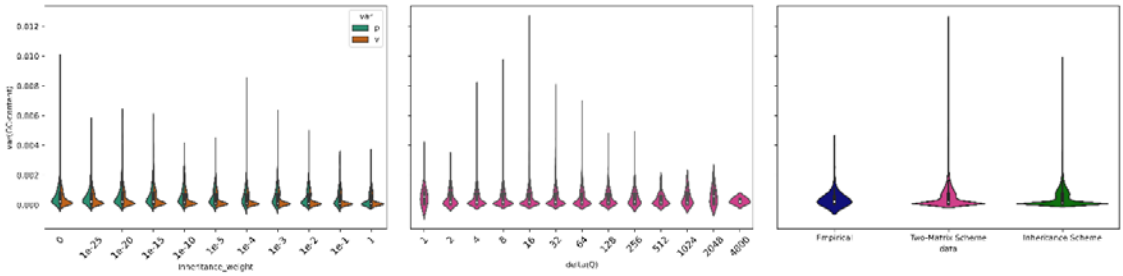

FIGURE 3| Variance in GC content in simulated and empirical datasets. The variance in GC content was calculated between sequences within each dataset.

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. 4, Appendix Figs. A.4-7). These metrics measure the difference between the inferred tree and the tree from which the alignment was simulated. If a 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 a stronger model violation through less homogeneity (for the rate matrix) and less stationarity (for the base frequencies). It is important to note that the results presented in Fig. 4 represent averages across datasets with varying numbers of taxa and sites. While Fig. A.5 represents the non-normalized RF metric, since the sub-figures all have the same number of taxa, nRF and RF are essentially the same.

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. 4, Appendix Fig. A.4). Finally, we see no correlation between the inheritance weights and having a significant p-value in a Kolmogorov-Smirnov test comparing the true and estimated branch lengths (slope = 0.52, $R^2 = 0.06$, p-value = 0.22), suggesting that violation of SRH assumptions in our evolutionary framework has no detectable effect on the estimation of branch lengths (Fig. 5, Appendix Fig. A.11).

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 Euclidean 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. 4, 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. 4).

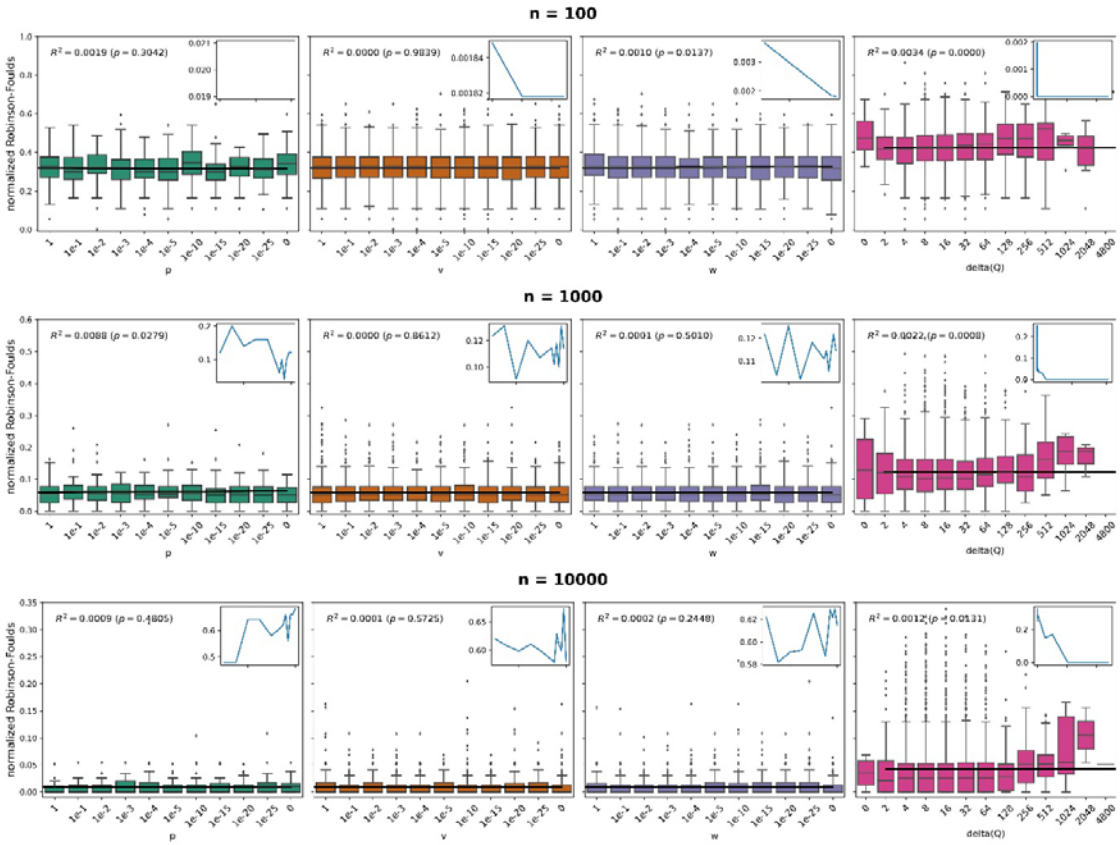

FIGURE 4. Normalized Robinson-Foulds distance between the estimated tree topology and the original tree topology as a function of the inheritance weight (ν, ω ,ρ) in the first simulation scheme, and the distance between the two matrices (delta(Q)) in the second simulation scheme, for different number of sites. The black line is the relevant linear regression. 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. of each inset plot matches that of the

bigger plot, labels have been excluded for clarity. 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 an inheritance 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 failing or passing the Kolmogorov-Smirnov test comparing the distributions of the true and estimated branch lengths (slope = 3.8E-4, $R^2 = 0.04$, p-value = 0.51), suggesting that violation of SRH assumptions in the convergent framework has limited effects on the estimation of branch lengths (Fig. 5, Appendix Fig. A.12, Supplementary material tables A.2).

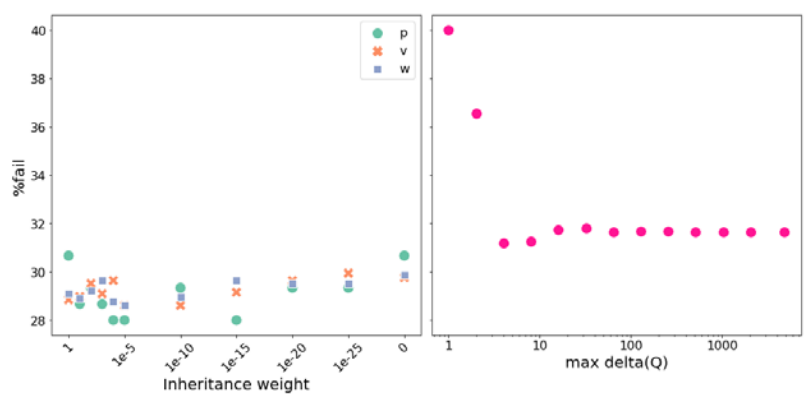

FIGURE 5. The percentage of datasets that pass the KS test (a dataset passes when there is no evidence to suggest that the inferred branch length distributions differ) as a function of the inheritance weights (ν,ρ,ω) (left-hand panel) and maximum Euclidean distance between the two matrices used to simulate the data (right-hand panel).

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. 2011a; Duchene, et al. 2017).

Surprisingly, our results show no correlation between errors in the topology or branch length distribution 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

topological inference that assume 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. Conant and Lewis 2001; Jermiin, et al. 2004; Duchene, et al. 2017).

It is noteworthy that our results from the different simulation schemes agree with the results from empirical data (Naser-Khdour, et al. 2019). When the model violations were more severe, like in the two-matrix scheme, we obtained a significant correlation between topological inference error and the extent of the violation of the SRH assumptions. In the inheritance scheme simulations, on the other hand, the violation of the SRH assumptions was milder, and therefore the phylogenetic inference was more robust. These results indicate that the extent of the violation of the SRH assumptions in the empirical datasets was higher than in our simulations under the inheritance scheme and it was probably closer to the extent of violation in the two-matrix scheme simulations. Moreover, they emphasize the impact of model violation due to non-SRH evolution on phylogenetic inference and suggest reducing model violation in phylogenetic analysis by using the new 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, the 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.

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-R 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.

Bird AP. 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. 8:1499-1504.

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.

Conant GC, Lewis PO. 2001. Effects of nucleotide composition bias on the success of the parsimony criterion in phylogenetic inference. Mol. Biol. Evol. 18:1024-1033.

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.

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.

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

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

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.

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.

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.

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

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, Gibb GC, Crimp EA, Penny D. 2010. Tinamous and moa flock together: mitochondrial genome sequence analysis reveals independent losses of flight among ratites. Syst. Biol. 59:90-107.

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.

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.

Yang Z, 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.

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.

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.

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.

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.