Nucleotide substitutions during speciation may explain substitution rate variation

Thijs Janzen[1*], Folmer Bokma[2], Rampal S. Etienne[1]

[1] Groningen Institute for Evolutionary Life Sciences, University of Groningen, Box 11103, 9700 CC Groningen, The Netherlands

[2] Center for Ecological and Evolutionary Synthesis (CEES) Department of BioSciences, University of Oslo, PO Box 1066, Blindern, 0316 Oslo, Norway

* corresponding author: t.janzen@rug.nl

RUNNING TITLE: Nucleotide substitutions during speciation

ABSTRACT

Although molecular mechanisms associated with the generation of mutations are highly conserved across taxa, there is widespread variation in mutation rates between evolutionary lineages. When phylogenies are reconstructed based on nucleotide sequences, such variation is typically accounted for by the assumption of a relaxed molecular clock, which is just a statistical distribution of mutation rates without much underlying biological mechanism. Here, we propose that variation in accumulated mutations may be partly explained by an elevated mutation rate during speciation. Using simulations, we show how shifting mutations from branches to speciation events impacts inference of branching times in phylogenetic reconstruction. Furthermore, the resulting nucleotide alignments are better described by a relaxed than by a strict molecular clock. Thus, elevated mutation rates during speciation potentially explain part of the variation in substitution rates that is observed across the tree of life.

Keywords: molecular clock, speciation, phylogenetic reconstruction, substitution rate variation

INTRODUCTION

Phenotypic diversification occurs at a higher rate in some clades than in others (Simpson 1945; van Valen 1985; Ricklefs 2006; Rabosky et al. 2007; Jansson and Davies 2008) and similarly, there is substantial variation across evolutionary lineages in the rate of molecular evolution (King and Wilson 1975), such as that of nucleotide sequences (Nabholz et al. 2008; Bromham 2011; Dowle et al. 2013; Sung et al. 2016). As a consequence, studies attempting to reconstruct the phylogeny of a clade often find that the sequence data do not support the assumption of a strict molecular clock, i.e. constant substitution rates across lineages. For such cases, phylogenetic inference software allows one to use a relaxed molecular clock (Drummond et al. 2006; Lepage et al. 2007), which assumes that the substitution rate varies between lineages according to a statistical distribution such as a gamma or lognormal distribution. However, the relaxed molecular clock thus introduces at least one additional degree of freedom, namely the variance of the distribution of substitution rates (although some argue that an uncorrelated relaxed clock in effect adds one additional degree of freedom *per branch* (Dornburg et al. 2012; Zhang and Drummond 2020; Douglas et al. 2021)). Moreover, the relaxed clock is a rather ad-hoc solution with little underlying biological reasoning (but see Lartillot and Poujol 2014; Lartillot et al. 2016; Saclier et al. 2018).

A first formal test to detect the impact of speciation on sequence evolution was formulated by Avise and Ayala (Avise and Ayala 1975, 1976), who distinguished gradual evolution from “punctuated equilibria” by comparing

64 sequence evolution in species-rich and species-poor clades. Whereas Avise and
65 Ayala found no evidence for increased sequence evolution in species-rich clades,
66 others did, in tetrapods (Mindell et al. 1989, 1990), sauropsids (Eo and DeWoody
67 2010) and angiosperms (Duchene and Bromham 2013; Bromham et al. 2015).
68 Furthermore, substitution rates have been found to be positively associated with
69 diversification rates (Fontanillas et al. 2007; Eo and DeWoody 2010; Lanfear et
70 al. 2010a, 2010b; Ezard et al. 2013, but see Goldie et al. 2011).

71 Several biological processes acting at speciation could lead to accelerated
72 sequence evolution, including, but not limited to founder effects, bottlenecks,
73 inbreeding, hybridization, selection for an increased mutation rate, divergent
74 selection and local adaptation (Venditti and Pagel 2010). Here, we explore how
75 such processes driving sequence evolution during speciation events might affect
76 phylogenetic reconstruction; we posit that differences in apparent substitution
77 rates between lineages are due to processes acting exclusively or predominantly
78 during speciation. Due to (effectively) random extinction of lineages, different
79 branches of a reconstructed phylogeny will differ in how often they experienced
80 such short episodes of accelerated substitution rates, resulting in differences in
81 apparent substitutions rates along these branches. Our approach is two-fold:
82 first, we explore whether inclusion of substitutions during speciation affects
83 phylogenetic inference, and, if so, which aspects of the inferred phylogenetic tree
84 are affected. Second, we explore whether substitutions during speciation can
85 explain variation in estimated substitution rates.

86

METHODS

We propose a model where substitutions accumulate not only along the branches of a phylogeny, but also at speciation events, including not only the internal nodes of the phylogeny but also those pruned from the phylogeny by extinction. We first make the standard assumption that gradual sequence evolution along a phylogenetic branch can be modeled as a time-homogeneous Markov process with substitution matrix:

$$\mathbf{Q} = \begin{bmatrix} -\mu_{AT}-\mu_{AC}-\mu_{AG} & \mu_{AT} & \mu_{AC} & \mu_{AG} \\ \mu_{TA} & -\mu_{TA}-\mu_{TC}-\mu_{TG} & \mu_{TC} & \mu_{TG} \\ \mu_{CA} & \mu_{CT} & -\mu_{CA}-\mu_{CT}-\mu_{CG} & \mu_{CG} \\ \mu_{GA} & \mu_{GT} & \mu_{GC} & -\mu_{GA}-\mu_{GT}-\mu_{GC} \end{bmatrix}$$

where μ_{ij} denotes the mutation rate from nucleotide i to nucleotide j. The transition probabilities of nucleotide substitutions after time t of gradual sequence evolution are then given by the matrix

$$\mathbf{P_a}(t) = \exp(\mathbf{Q}t)$$

where the subscript **a** indicates anagenetic change, i.e. gradual accumulation of substitutions over time. This matrix can be multiplied with an initial probability vector at time $t = 0$ to yield the probabilities for each of the four nucleotides at time t.

In addition to gradual sequence evolution over time, we assume that sequences may change rapidly during speciation. We can thus assume another matrix $\mathbf{P_c}$ (subscript c for "cladogenetic") that describes the nucleotide transition probabilities during a single speciation event.

The processes that may accelerate sequence evolution during speciation, such as founder effects, bottlenecks, inbreeding, hybridization, and adaptation to novel environments, may well result in different kinds of substitutions than those that take place over time in established species. However, for mathematical convenience we will here assume that we can write:

$$\mathbf{P}_{\mathrm{c}} = \exp\left(\mathbf{Q}\tau\right)$$

where τ is a parameter that measures the effect of substitutions during speciation. In other words, we assume that nucleotide sequence evolution is only accelerated during speciation events, but not qualitatively altered: the μ_{ij} used in $\mathbf{P}_{\mathrm{c}}$ must be identical to those used in $\mathbf{P}_{\mathbf{a}}$. The acceleration is then measured by parameter τ: a single speciation event causes as much sequence evolution as τ years of gradual evolution over time within each lineage. Thus, larger values of τ correspond to a larger experienced effect at the nodes, similar to sequence evolution along a branch of length τ. For $\tau = 0$, $\mathbf{P}_{\mathrm{c}}$ becomes the identity matrix, and our model reduces to the standard model of sequence evolution that only assumes substitutions along phylogenetic branches. Important to note here is that both daughter lineages resulting from a speciation event experience substitutions independently (see the Supplement for a model where substitutions in both daughter lineages are dependent on each other). Furthermore, we emphasize that we assume the speciation process to happen in a similar fashion across a tree, assuming an identical τ for all nodes in the tree. Later versions of the model could potentially relax this assumption, provided independent information about speciation dynamics.

For simplicity we assume in our simulations that sequence evolution can be modeled as a Jukes-Cantor process (Jukes and Cantor 1969), for which **Q** is given by:

$$\mathbf{Q} = \begin{pmatrix} -\frac{3\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & -\frac{3\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & -\frac{3\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & -\frac{3\mu}{4} \end{pmatrix}$$

Several existing software packages (e.g. the R package phangorn, Schliep 2011; the python module pyvolve, Sipos *et al.* 2011; the R package phylosim, Spielman & Wilke 2015) provide algorithms to simulate sequence evolution along the branches of the phylogeny, given a rooted phylogeny and a root sequence (e.g. some arbitrary sequence assumed to represent the ancestral sequence), by applying the transition matrix sequentially along the phylogenetic tree. Here, we extend this methodology to also include substitutions accumulated at the nodes of the phylogeny. We implemented this in the R package 'nodeSub' (https://github.com/thijsjanzen/nodeSub, and soon available on CRAN).

143 *Testing the impact of node substitution models using simulations*

144 To identify the amount of error in phylogenetic inference caused by
145 assuming a (relaxed) molecular clock when substitutions actually arise (in part)
146 during speciation, we simulated sequence evolution on known trees and then
147 reconstructed the phylogeny from the simulated sequences, assuming strict and
148 relaxed molecular clocks. We simulated sequence evolution with the node
149 substitution model introduced above, with various degrees of sequence
150 accumulation at the nodes of the tree (τ), and with various extinction rates. We
151 then compared the resulting trees with the original true tree using a number of
152 statistics: the gamma statistic (Pybus and Harvey 2000), the beta statistic
153 (Aldous 2001), the mean branch length (Faith 1992; Clarke and Warwick 2001),
154 crown age, the normalized Lineages Through Time (nLTT) statistic (Janzen et al.
155 2015) and the Jenson-Shannon distance metric comparing the Laplacian
156 spectrum (Lewitus and Morlon 2016).

157 Phylogenetic reconstruction was performed with BEAST2 (Bouckaert et
158 al. 2019) using the R package *babette* (Bilderbeek and Etienne 2018). BEAST2
159 inference was performed using default priors, with a birth-death prior as tree
160 prior (or a Yule prior if the extinction rate was zero), the Jukes-Cantor nucleotide
161 substitution model, and a strict or relaxed clock model. The BEAST chain was run
162 for 10 million steps, whilst sampling a tree every 5000 steps. After completion,
163 the first 10% of the chain was discarded as burn-in.

164

165 *Assessing error in phylogenetic reconstruction: the twin tree*

166 Errors observed when comparing with the true tree include both errors
167 incurred by the node substitution model chosen, and errors accumulated in the

168 phylogenetic inference process even when the models used in inference are
169 identical to those generating the data (e.g. stochasticity in substitution
170 accumulation, stochasticity in phylogenetic tree creation). Furthermore,
171 additional effects arising during alignment simulation might affect our findings,
172 such as the impact of parameter values (sequence length, substitution rate, birth
173 rate, death rate), and of multiple substitutions at the same site (the node-
174 density-effect) as well as potential biases or interactions between summary
175 statistics. To correct for these effects, so as to isolate the error induced by using a
176 node substitution model from other sources of error, we inferred a phylogenetic
177 tree for a twin alignment (*sensu* Bilderbeek, Laudanno & Etienne 2020). This
178 twin alignment has exactly the same number of substitutions and is based on the
179 same true tree, but instead of using a node substitution model to generate the
180 alignment, it results from using either a strict-clock or relaxed-clock substitution
181 model. Using this twin alignment, we performed phylogenetic reconstruction
182 with BEAST2 as for the original alignment, and estimated the same summary
183 statistics for the posterior distribution of trees. The error introduced by the node
184 substitution model is then the difference between the error of the node
185 substitution posterior and the error in the twin posterior. In summary, we use
186 this twin approach as a control treatment, in order to correct for all potential
187 sources of additional error other than that of our proposed substitution model.
188
189 *Obtaining a twin alignment*
190 We generated a twin alignment conditional on a phylogeny, a node
191 substitution model, and a mutation rate. Because an alignment generated using a
192 node substitution model (with $\tau > 0$) has accumulated substitutions at the nodes

in addition to those along the branches, the overall number of substitutions accumulated is higher than for an alignment simulated using the same mutation rate and a model with substitutions only on the branches. Thus, in order to generate a *twin* alignment that contains the same amount of information (substitutions) we increased the mutation rate. We did this by calculating the estimated time spent at the nodes, relative to the time spent on the branches, and using this as an estimate of the expected fraction of the number of substitutions on the nodes, relative to the number of substitutions on the branches, assuming that substitutions accumulate at the same rate on both branches and nodes. That is, the mutation rate used in generating the twin alignment is calculated as:

$$\mu_{twin} = \mu\left(1 + \frac{\tau(2N + H)}{\sum t_{branch}}\right) \quad (1)$$

where μ is the mutation rate used in the node substitution model, τ is the time spent on the node, N is the number of internal nodes in the tree, H is the number of hidden nodes in the tree and $\sum t_{branch}$ is the total branch length of the tree. The factor $2N$ arises from the independent accumulation of substitutions during a node substitution event for both daughter lineages.

During simulation of node substitution alignments, we kept track of the substitutions accumulated at each node and branch, which allowed us to directly measure the contribution of substitutions accumulated at the nodes (i.e. $\tau(2N + H)$) relative to those accumulated at the branches (i.e.. $\sum t_{branch}$) This provided us with an estimate of μ_{twin}, and with an estimate of the total number of substitutions arising during simulation of the alignment. We then used the obtained estimate for μ_{twin} to generate <u>*twin*</u> alignments, again tracking all

substitutions, until we obtained an alignment exactly matching the number of accumulated substitutions of the alignment simulated with the node substitution model.

Node-density-effect

The method we used to simulate substitutions along branches (and nodes) ignores repeated mutations at the same site, which may lead to a node-density-effect. Because the node-density-effect can mask the effect of node substitutions, we made sure in two distinct ways that our results are not affected by this effect. Firstly, by using a *twin* alignment, any resulting node-density-effects are mirrored in the *twin* alignment as well, ensuring that any additional errors picked up do not reflect errors induced by the node-density-effect. Secondly, we repeated our analysis using a different simulation method that explicitly tracks repeated mutations for the Jukes-Cantor model (see Supplementary Information for details and results). We find that this more explicit simulation method yielded virtually identical results to the more general approach described in the main text.

Simulation settings

We generated birth-death trees with varying degrees of extinction rate d in [0, 0.1, 0.3, 0.5] and a single speciation rate of b = 1. Trees were simulated conditional on 100 tips, using the function `sim.bd.taxa` from the R package TreeSim (Stadler 2011). Across all settings, we simulated sequences of 10 kb, with μ = 0.001.

Varying the time spent on the nodes relative to the crown age

We varied τ in [0, 0.01, 0.05, 0.1, 0.2, 0.4] times the crown age (e.g. when the crown age of the simulated tree is 3MY, τ = [0, 0.03, 0.15, 0.3, 0.6, 1.2] MY). Again, for each combination of τ and extinction (d = [0, 0.1, 0.3, 0.5]) we simulated 100 trees and for each tree we generated one node substitution alignment and one *twin* alignment.

The impact of tree balance

In unbalanced trees, some terminal branches are connected by many more past branching events to the root of the tree than are other terminal branches. Hence, we expect that balance of a tree might have a substantial effect on the error in phylogenetic inference: less balanced trees are expected to have higher error. To test this, we compared fully balanced (= 10.0) with extremely unbalanced "caterpillar" trees (= -2). We did so by simulating the branching times of a birth-death tree, and assigning these to a fully balanced or fully unbalanced topology. Thus, the only difference between the trees is the topology. Then, for both the balanced and unbalanced tree a node substitution alignment was generated, with the same number of total substitutions, and setting τ as a function of crown age.

As an extra check, we also generated a node substitution alignment for the original birth-death tree from which the branching times were used. For all three alignments we inferred a phylogenetic tree as in the other scenarios, and compared the error in phylogenetic inference. For caterpillar trees with extreme unbalance we were unable to calculate the Laplacian Spectrum, hence we omitted the Laplacian Spectrum summary statistic in this analysis.

266 *Support for strict and relaxed clock models*

267 To test whether the alignment originating from a process with node
268 substitutions was better described by a relaxed than by a fixed clock model, we
269 repeated the analysis, but now inferred the marginal likelihood of the relaxed
270 clock and strict clock models using the "NS" package for BEAST2, which applies
271 Nested Sampling to obtain the marginal posterior likelihood for both models
272 (Russel et al. 2019). We used the function '`bbt_run`' from the *babette* package
273 (Bilderbeek and Etienne 2018) in combination with the function
274 '`create_ns_mcmc`' from the *beautier* package (Bilderbeek and Etienne 2018).
275 This performs a Nested Sampling MCMC run using BEAST2, which runs until
276 convergence is detected. Then, we converted the obtained marginal likelihoods
277 to a relative weight for each model (by dividing both marginal likelihoods by
278 their sum) which allows for comparison of posterior support for each model
279 across parameter settings and trees.

280

RESULTS

Summary statistics

We compared summary statistics of trees inferred from alignments using the node substitution model, with summary statistics of twin trees inferred from alignments with identical information content, but generated without the node substitution model (e.g. with only substitutions along the branches, and a fixed clock rate). We find that summary statistics that are influenced by branching times are affected (Figure 1, gamma, nLTT statistic, mean branch length and crown age). For these summary statistics, we find an increased difference as the time spent on the nodes τ increases. The differences in the values of these summary statistics manifest themselves across all extinction rates, although the effect tends to be somewhat smaller for higher extinction rates.

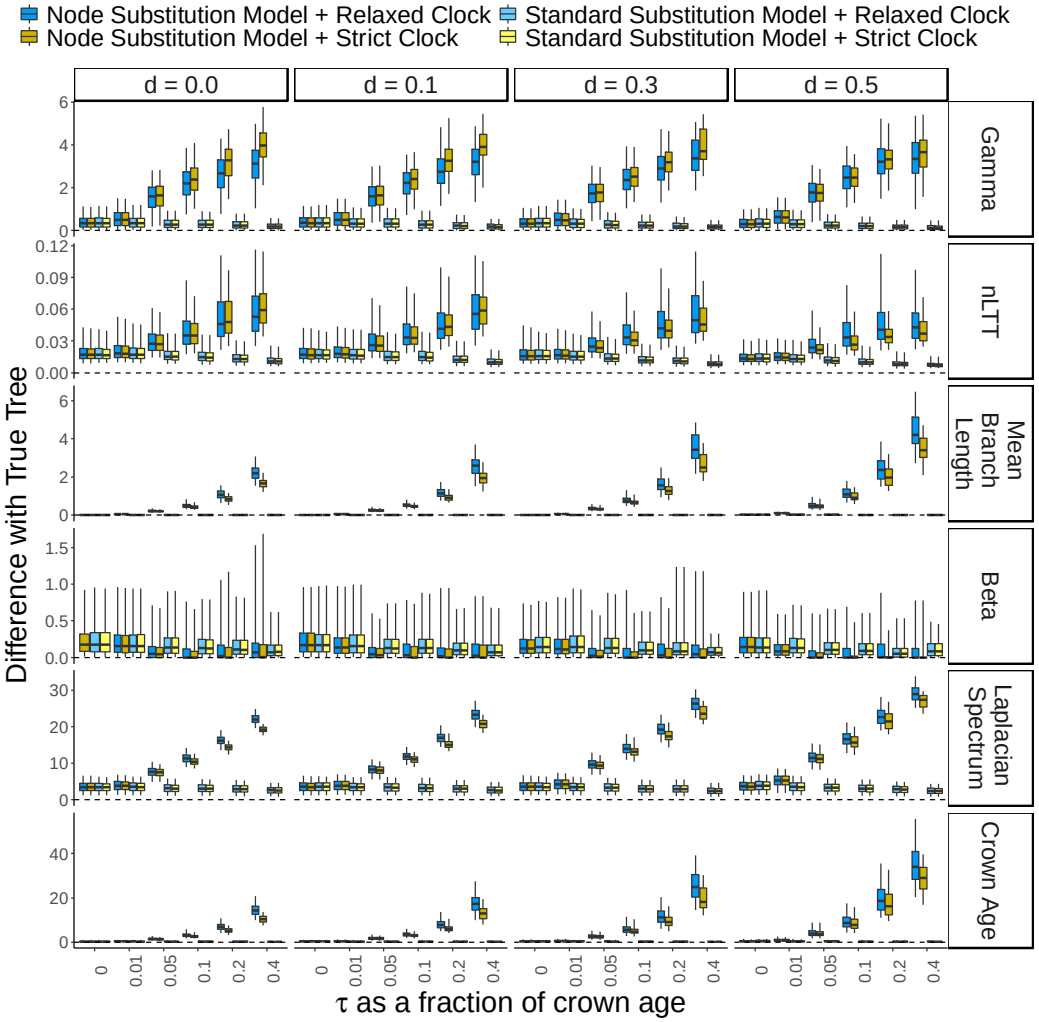

Figure 1. Difference in summary statistic values for trees inferred from an alignment generated with node substitutions, and *twin* trees that were inferred from an alignment generated without node substitutions (grey boxplots), both compared with the summary statistics of the true tree. We explored τ (the amount of time spent on each node) as a fraction of crown age (horizontal axis), and the impact of extinction (d, columns). The summary statistics are the beta and gamma statistic, Laplacian spectrum, mean branch length, nLTT statistic, and crown age. The figure shows that with increasing τ , trees inferred from an alignment generated with node substitutions show larger differences with the true tree than trees inferred from an alignment generated without node substititions. Differences with the true tree are larger for trees inferred using the strict clock model than for those using the relaxed clock model, but only for the alignment generated with node substitutions.

The impact of tree balance

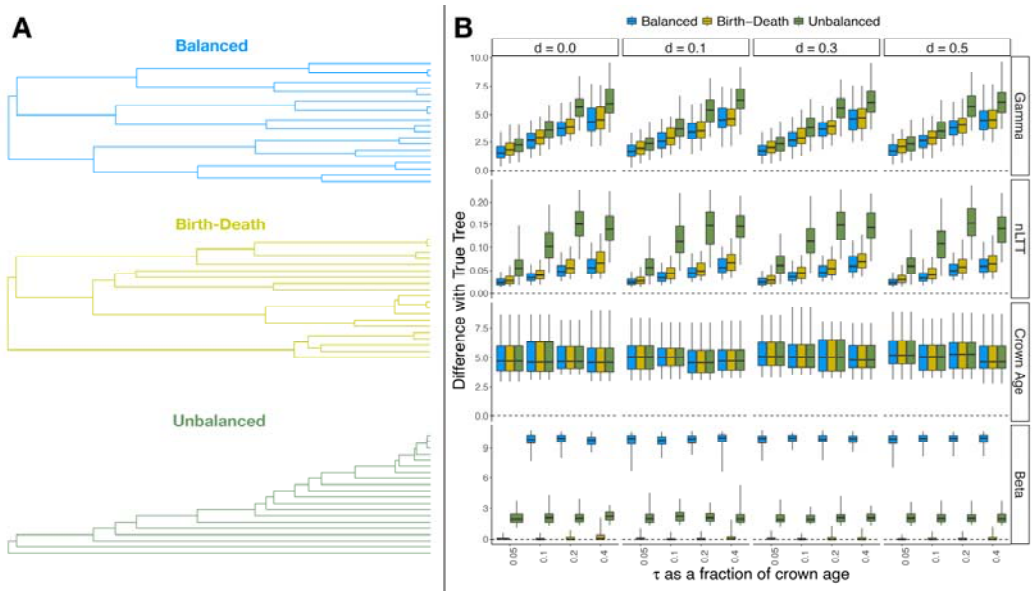

Figure 2. Effect of the node substitution model for phylogenies differing in balance **A**) example plots of a randomly generated birth-death tree (top), a fully balanced tree generated using the same branching times as the birth-death tree (middle) and a very unbalanced tree generated using the same branching times as the birth-death tree (bottom). Shown are trees with 20 tips for illustrative purposes, but results in **B** are from trees with 100 tips. **B)** Difference in summary statistic with the true birth-death tree for phylogenetic trees inferred from alignments generated using the node substitution model on either balanced, unbalanced or random trees. We explore τ as a fraction of crown age (horizontal axis), and the impact of extinction (d, columns). The dotted line indicates zero difference with the true tree. The summary statistics are the beta and gamma statistic, nLTT statistic and tree height. Balanced and birth-death trees tend to have similar inferred error, whereas unbalanced trees differ strongly, with a much larger error for the gamma and nLTT statistic.

Tree balance clearly influences the sensitivity of inference to node substitutions (Figure 2). The inference error is larger for unbalanced trees, again only for the gamma and the nLTT statistic. Fully balanced trees show slightly less error than birth-death trees. Overall, all three types of trees show an increased error when alignments are generated using the node substitution model. Errors are particularly large for the beta statistic, but that is expected because it measures topological features of the tree that we modified artificially.

335 *Support for strict and relaxed clock models*

336 We compared the relative support for each model, reflected by the
337 relative weight of the marginal likelihood. With an increasing amount of time
338 spent at the nodes τ, the median weight of the relaxed clock model increases for
339 the node substitution alignment, with generally (across extinction rates) a higher
340 weight than the strict clock model for values of τ that are equal or larger than 0.1
341 times the crown age (Figure 3). For the twin alignment, the strict clock model is
342 preferred, as expected, because this is the generating model.

343 For low values of τ (smaller than 0.1 times the crown age), we do not find any
344 effect of the balance of the tree on the marginal likelihood of the relaxed clock
345 model (Figure 4), in line with our finding above. However, for intermediate
346 values of τ (0.1 and 0.2), we find that unbalanced trees tend to have a higher
347 marginal weight for the relaxed clock model. For high values of τ (0.4), we find
348 that the marginal weight for the relaxed clock model is always higher, regardless
349 of the balance of the tree.

350

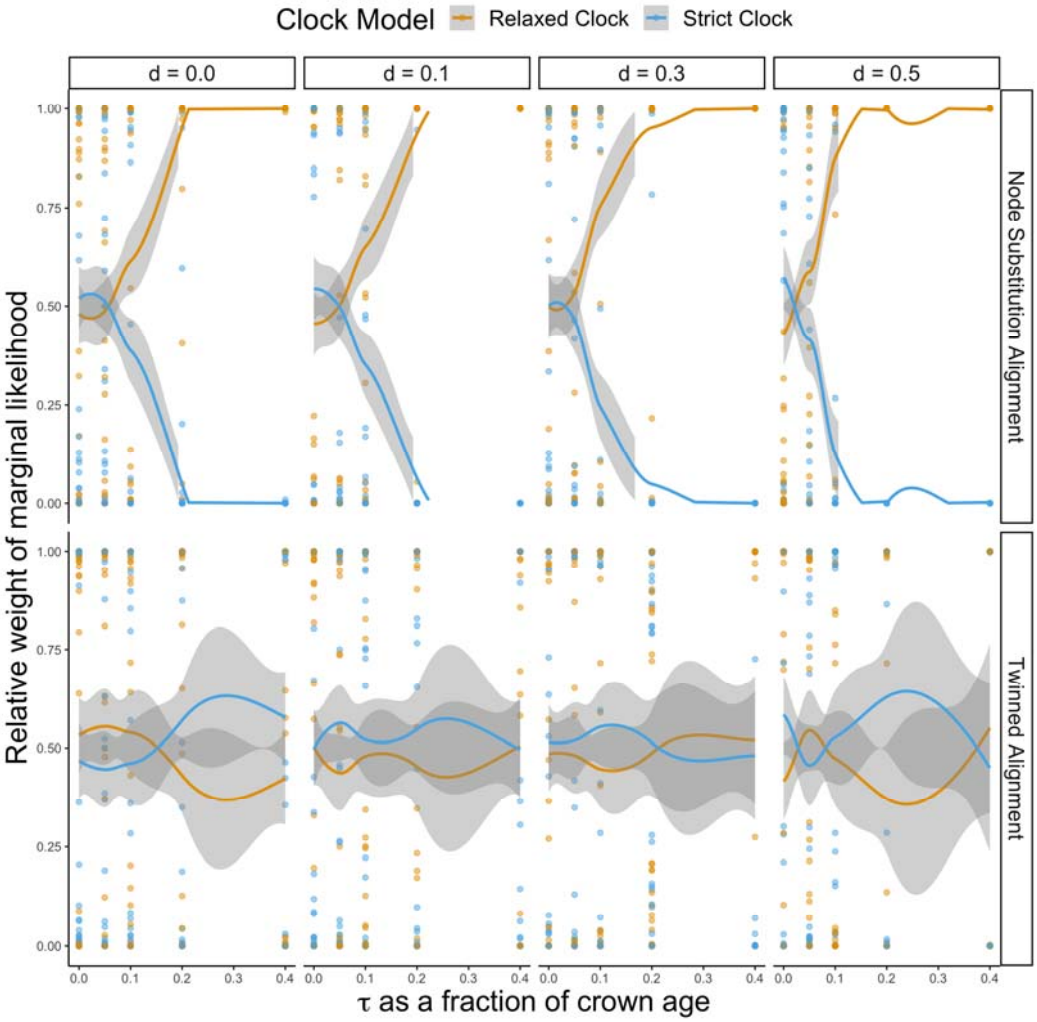

Figure 3. Marginal likelihood weight of the relaxed and strict clock model for varying time spent on the nodes (τ), where τ is chosen as fraction of the crown age. Alignments generated with a node substitution model (top row) are compared with alignments generated without node substitutions (bottom row). Per parameter combination, 100 replicate trees were analyzed. Because many dots are plotted on top of each other, we use solid lines to indicate the best fitting locally estimated scatterplot smoothing (LOESS), and the 95% Confidence interval (grey shaded area) of the LOESS curve. As the time spent on the nodes increases, posterior support for the relaxed clock model increases, but only if the alignment was generated with a node substitution model.

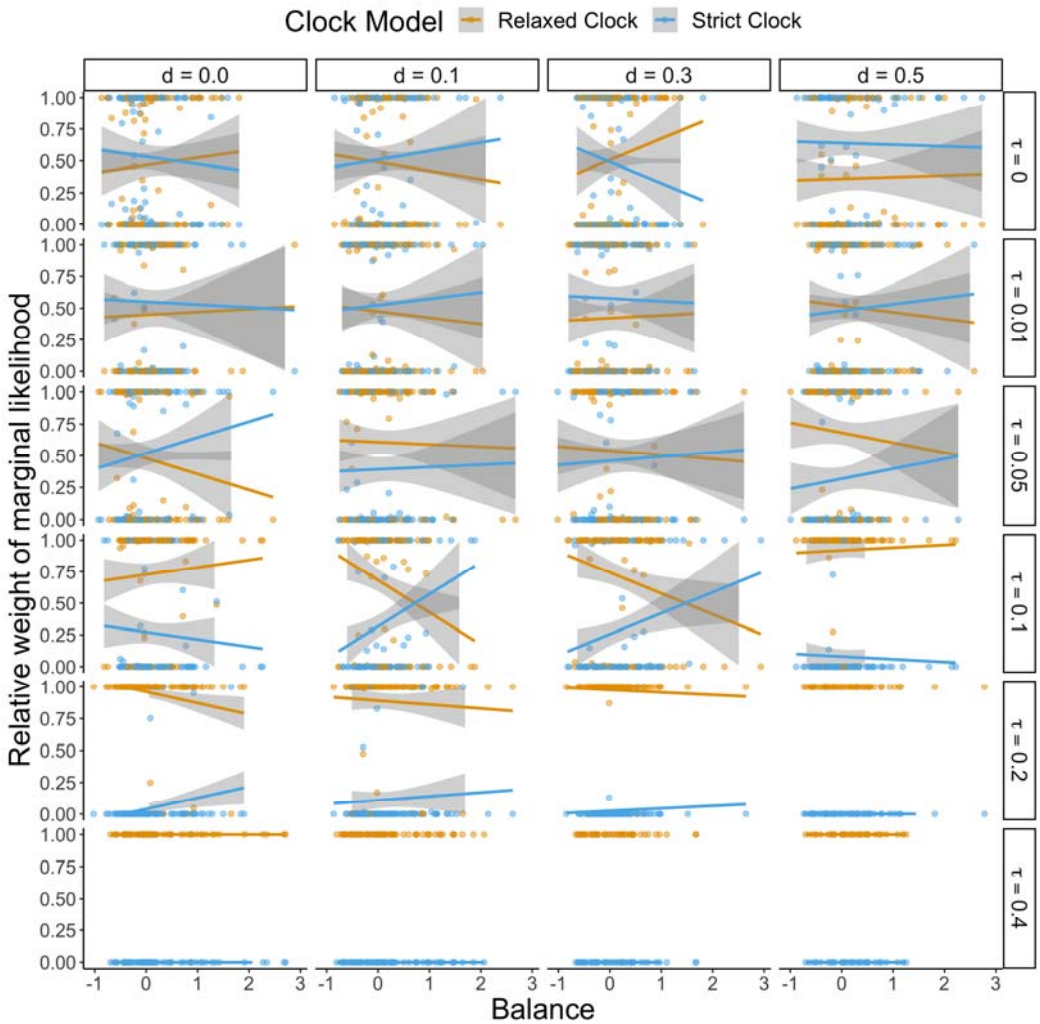

Figure 4. Marginal likelihood weight of the relaxed clock model for trees of varying balance, split out across different extinction rates (d = [0, 0.1, 0.3, 0.5]) and time spent on the nodes (τ), where τ is chosen as fraction of the crown age (e.g. τ = 0.1 reflects a node time of 10% of the crown age). Per parameter combination, Solid lines indicate the best fitting linear regression and the 95% confidence interval (grey shaded area) of regression. With increasing values of τ, the relative weight of the relaxed clock model becomes larger. For smaller values of τ, the relative weight of the relaxed clock model is negatively correlated with the balance of the tree, with unbalanced trees having a higher relative weight.

377 DISCUSSION

378 We have shown that an increased substitution rate during speciation
379 events potentially provides a mechanistic explanation of variation in substitution
380 rates across the branches of phylogenetic trees. Trees inferred from alignments
381 generated with this substitution model differ substantially from trees inferred
382 from alignments generated with a standard substitution model, especially
383 concerning branching times. Furthermore, we find that this new substitution
384 model can potentially explain widespread support for relaxed molecular clocks.

385 If sequence evolution mainly occurs during speciation, this would lead to
386 a correlation between species richness and substitution rate. However, this
387 correlation could also be an artifact of phylogenetic reconstruction known as the
388 node-density-effect (Fitch and Bruschi 1987; Fitch and Beintema 1990). The
389 node-density-effect reflects the inability to detect multiple mutations occurring
390 at the same site, thus causing an underestimate of the true branch length,
391 especially for longer branches where the probability of multiple mutations
392 occurring at the same site is higher. Because species-rich parts of phylogenies
393 tend to have shorter branches, sequence evolution in these species-rich parts is
394 less underestimated than in species-poor parts, causing a correlation between
395 the number of observed substitutions and species diversity. Pagel et al. tested for
396 the impact of speciation events, and of the node-density-effect in 122
397 phylogenies, spanning 4 taxa (Pagel et al. 2006). Using previously demonstrated
398 methodology to detect the node-density effect (Webster et al. 2003; Venditti et
399 al. 2006), they showed that in 57 of the examined phylogenies, they could detect
400 a signature of increased sequence evolution during speciation events. However,
401 this was the result of the node-density effect in 22 out of these 57 phylogenies.

402 Here, disentangling sequence evolution during speciation from confounding
403 factors such as the node-density effect, but also stochasticity in tree simulation,
404 stochasticity during alignment simulation and error or bias in tree inference, has
405 proven to be a non-trivial endeavor. In order to assess the impact of node
406 substitutions, we therefore separated error due to assuming an alternative
407 substitution model from the errors introduced by the factors mentioned above.
408 To do so we extended the twinning approach (introduced by Bilderbeek,
409 Laudanno and Etienne (2020)) to assess the impact of choosing a different tree
410 prior) to explore the impact of a different substitution model. The twinning
411 approach succeeds by replicating the chosen analysis pipeline, but using *control*
412 data that have been generated using known models and priors. The impact of the
413 node substitution model then follows from the difference between results
414 obtained with the node substitution model and results obtained with the *twin*
415 (control) pipeline: errors are then due to model misspecification, and not
416 stochastic uncertainty produced by the analysis pipeline.

417 One might expect that a high extinction rate, by elevating numbers of
418 hidden nodes, would lead to a greater impact of node substitutions. It may
419 therefore be counterintuitive that in our simulation study we did not find such
420 an effect of higher rates of extinction. However, we conditioned our alignments
421 on the same total number of substitutions, to ensure that alignments with and
422 without node substitutions contained the same information content. Thus, with
423 higher extinction and hence more hidden nodes, fewer substitutions occur on the
424 observed nodes. Because the number of hidden nodes is proportional to branch
425 length (Eq.2), it is interpreted as substitutions on the branches. Consequently,
426 higher extinction rates erase the signature of node substitutions. Potentially, this

427 provides a way to distinguish between phylogenetic models: although every
428 constant-rate birth-death model has a corresponding zero-extinction model with
429 a time-varying speciation rate that yields the same probability of the
430 reconstructed tree (Nee et al. 1994; Louca and Pennell 2020), the resulting
431 alignments under the node substitution model will not be similar. Because the
432 birth-death tree includes extinction events, substitution patterns will be
433 different from those of the tree generated with the time-varying speciation rate
434 model.

435 Distinguishing phylogenetic models will become more feasible if some of
436 the simplifying assumptions made here are relaxed. The model we propose here
437 takes the simplest form, assuming a Jukes-Cantor (Jukes and Cantor 1969)
438 substitution matrix, identical substitution rates, identical substitution matrices
439 between nodes and branches, and constant birth-death rates over time. These
440 assumptions were made as a most basic starting point, but can be relaxed in
441 future analyses, for instance by introducing a different substitution matrix at the
442 nodes, or by studying the effect of node substitutions on trees that are generated
443 by diversity-dependent speciation rates (Etienne et al. 2012). By starting with
444 the most tractable version of the node substitution model, we have provided a
445 first proof of concept of the potential impact of node substitutions without
446 overcomplicating matters.

447 Previous methods have applied rather ad-hoc corrections to account for
448 differences in substitution rates across different branches in the same
449 phylogeny, typically referred to as the 'relaxed clock' approach. These methods
450 provide satisfying statistical solutions to account for variation in substitution
451 rates, but refrain from providing biological explanations for this observed

452 phenomenon. The node substitution model we introduce here provides this
453 explanation: branches that have accumulated a number of 'hidden' branching
454 events, e.g. speciation events of species that have subsequently gone extinct,
455 have a higher number of accumulated substitution events during these 'hidden'
456 speciation events. When we compared the marginal likelihood of the relaxed
457 clock model versus the strict clock model for alignments generated with the
458 node substitution model, we found that marginal likelihoods for the relaxed
459 clock model are much higher. This indicates that our proposed process of
460 accumulating substitutions during speciation events can generate patterns in the
461 alignment that are picked up by phylogenetic methods as evidence for a relaxed
462 clock model, without actually using a relaxed clock model.

463 The notion of accelerated evolution during speciation events ties
464 in closely with the theory of punctuated equilibrium; where Eldredge and Gould
465 (1972) proposed that evolution perhaps is not a gradual process, but rather a
466 process with distinct bursts of phenotypic and morphological change. Their
467 theory was influenced by ideas like Lerner's "genetic homeostasis" (Lerner
468 1954), which had earlier inspired Mayr (1954) to suggest that the formation of
469 new species involves "genetic revolutions". Our framework provides a step
470 towards being able to test this notion, where information on the estimated
471 fraction of substitutions accumulated at the nodes can directly inform us about
472 whether the majority of substitutions is accumulated over long periods of time in
473 established lineages (e.g. along branches), or during speciation (e.g. at the
474 nodes).

475 The present study aims to demonstrate that substitutions accumulated
476 during speciation might explain the prevalence of the relaxed molecular clock in

477 phylogenetic analysis. We found that substitutions during speciation may
478 profoundly affect phylogenetic inference: if node substitutions are not taken into
479 account, branching times tend to be overestimated, even when a relaxed clock is
480 used to counteract the effect of "hidden nodes". This suggests that incorporation
481 of a node substitution model improves phylogenetic inference. However, the
482 computation of the likelihood of our model (and estimation of associated τ
483 values), is non-trivial, and might escape computational feasibility due to the high
484 state space required. Manceau *et al.* (2020) have taken a first step towards
485 formulating such a likelihood. They introduced an alternative solution for
486 punctuated equilibrium-like patterns in molecular evolution through the
487 implementation of spikes of substitution, e.g. moments in time at which there is
488 an increased substitution rate. They let these moments occur at speciation
489 events, and also model the probability of such an event happening at a speciation
490 event (rather than assuming that they always occur, like we did here). However,
491 they have to assume both the topology and branching times to be fixed. This
492 provides a promising first step, but given that the number of hidden nodes
493 depends on the diversification rates, what is ultimately needed is a new method
494 that jointly infers the diversification model and the substitution model.

495 With our introduction of the node substitution model, we hope to stimulate
496 discussion on the biological explanation of variation in substitution rates within
497 and across phylogenies. Furthermore, we hope to have set a first step in
498 improving our understanding of this variation, and improving phylogenetic
499 inference as a whole.

500

501 CONFLICT OF INTEREST

The authors have no conflicts of interest to report.

ACKNOWLEDGEMENTS

We thank Sebastian Hoehna and three anonymous referees for providing comments on the manuscript. FB was supported by the Research Council of Norway, grant 263149. RSE thanks the Netherlands Organization for Scientific Research (NWO) for financial support through VIDI and VICI grants. TJ thanks the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster.

514 REFERENCES

515 Aldous D. 2001. Stochastic models and descriptive statistics for phylogenetic
516 trees, from Yule to today. Stat. Sci. 16:23–34.

517 Avise J.C., Ayala F.J. 1975. Genetic change and rates of cladogenesis. Genetics.
518 81:757–773.

519 Avise J.C., Ayala F.J. 1976. Genetic Differentiation in Speciose Versus Depauperate
520 Phylads: Evidence from the California Minnows. Evolution. 30:46.

521 Bilderbeek R.J.C., Etienne R.S. 2018. babette: BEAUti 2 , BEAST2 and Tracer for
522 R. 2018:2034–2040.

523 Bilderbeek R.J.C., Laudanno G., Etienne R.S. 2020. Quantifying the impact of an
524 inference model in Bayesian phylogenetics. Methods Ecol. Evol.:1–8.

525 Bouckaert R., Vaughan T.G., Barido-Sottani J., Duchene S., Fourment M.,
526 Gavryuskina A., Heled J., Jones G., Kuhnert D., De Maio N., Matschiner M., K.
527 Mendes F., Muller N.F., Ogilvie H.A., du Plessis L., Popinga A., Rambaut A.,
528 Rasmussen D., Siveroni I., Suchard M.A., Wu C.H., Xie D., Zhang C., Stadler T.,
529 Drummond A.J. 2019. BEAST 2 .5: An advanced software platform for
530 Bayesian evolutionary analysis. PLoS Comput. Biol. 15:e1006650.

531 Bromham L. 2011. The genome as a life-history character: Why rate of molecular
532 evolution varies between mammal species. Philos. Trans. R. Soc. B Biol. Sci.
533 366:2503–2513.

534 Bromham L., Hua X., Lanfear R., Cowman P.F. 2015. Exploring the relationships
535 between mutation rates, life history, genome size, environment, and species
536 richness in flowering plants. Am. Nat. 185:508–524.

537 Clarke K., Warwick R. 2001. A further biodiversity index applicable to species
538 lists: variation in taxonomic distinctness. Mar. Ecol. Prog. Ser. 216:265–278.

539 Dornburg A., Brandley M.C., McGowen M.R., Near T.J. 2012. Relaxed clocks and
540 inferences of heterogeneous patterns of nucleotide substitution and
541 divergence time estimates across whales and dolphins (Mammalia:
542 Cetacea). Mol. Biol. Evol. 29:721–736.
543 Douglas J., Zhang R., Bouckaert R. 2021. Adaptive dating and fast proposals:
544 Revisiting the phylogenetic relaxed clock model. PLOS Comput. Biol.
545 17:e1008322.
546 Dowle E.J., Morgan-Richards M., Trewick S.A. 2013. Molecular evolution and the
547 latitudinal biodiversity gradient. Heredity (Edinb). 110:501–510.
548 Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics
549 and dating with confidence. PLoS Biol. 4:e88.
550 Duchene D., Bromham L. 2013. Rates of molecular evolution and diversification
551 in plants: Chloroplast substitution rates correlate with species-richness in
552 the Proteaceae. BMC Evol. Biol. 13:1.
553 Eldredge N., Gould S.J. 1972. Punctuated equilibria: an alternative to phyletic
554 gradualism. Models In Paleobiology. Freeman Cooper and Co. p. 82–115.
555 Eo S.H., DeWoody J.A. 2010. Evolutionary rates of mitochondrial genomes
556 correspond to diversification rates and to contemporary species richness in
557 birds and reptiles. Proc. R. Soc. B Biol. Sci. 277:3587–3592.
558 Etienne R.S., Haegeman B., Stadler T., Aze T., Pearson P.N., Purvis A., Phillimore
559 A.B. 2012. Diversity-dependence brings molecular phylogenies closer to
560 agreement with the fossil record. Proc. R. Soc. B Biol. Sci. 279:1300–1309.
561 Ezard T.H.G., Thomas G.H., Purvis A. 2013. Inclusion of a near-complete fossil
562 record reveals speciation-related molecular evolution. Methods Ecol. Evol.
563 4:745–753.

564 Faith D.P. 1992. Conservation evaluation and phylogenetic diversity. Biol.
565 Conserv. 61:1–10.
566 Fitch W.M., Beintema J.J. 1990. Correcting parsimonious trees for unseen
567 nucleotide substitutions: The effect of dense branching as exemplified by
568 ribonuclease. Mol. Biol. Evol. 7:438–443.
569 Fitch W.M., Bruschi M. 1987. The evolution of prokaryotic ferredoxins--with a
570 general method correcting for unobserved substitutions in less branched
571 lineages. Mol. Biol. Evol. 4:381–394.
572 Fontanillas E., Welch J.J., Thomas J.A., Bromham L. 2007. The influence of body
573 size and net diversification rate on molecular evolution during the radiation
574 of animal phyla. BMC Evol. Biol. 7:1–12.
575 Goldie X., Lanfear R., Bromham L. 2011. Diversification and the rate of molecular
576 evolution: No evidence of a link in mammals. BMC Evol. Biol. 11:286.
577 Jansson R., Davies T.J. 2008. Global variation in diversification rates of flowering
578 plants: Energy vs. climate change. Ecol. Lett. 11:173–183.
579 Janzen T., Höhna S., Etienne R.S.R.S. 2015. Approximate Bayesian Computation of
580 diversification rates from molecular phylogenies: introducing a new
581 efficient summary statistic, the nLTT. Methods Ecol. Evol. 6:566–575.
582 Jukes T., Cantor C. 1969. Evolution of protein molecules. Mamm. Protein Metab.
583 21.
584 King M.C., Wilson A.C. 1975. Evolution at two levels in humans and chimpanzees.
585 Science. 188:107–116.
586 Lanfear R., Ho S.Y.W., Love D., Bromham L. 2010a. Mutation rate is linked to
587 diversification in birds. Proc. Natl. Acad. Sci. U. S. A. 107:20423–20428.
588 Lanfear R., Welch J.J., Bromham L. 2010b. Watching the clock: Studying variation

589 in rates of molecular evolution between species. Trends Ecol. Evol. 25:495–
590 503.
591 Lartillot N., Phillips M.J., Ronquist F. 2016. A mixed relaxed clock model. Philos.
592 Trans. R. Soc. B Biol. Sci. 371.
593 Lartillot N., Poujol R. 2014. Correlated evolution of substitution rates and
594 quantitative traits. .
595 Lepage T., Bryant D., Philippe H., Lartillot N. 2007. A general comparison of
596 relaxed molecular clock models. Mol. Biol. Evol. 24:2669–80.
597 Lerner I.M. 1954. Genetic Homeostasis. .
598 Lewitus E., Morlon H. 2016. Characterizing and comparing phylogenies from
599 their laplacian spectrum. Syst. Biol. 65:495–507.
600 Louca S., Pennell M.W. 2020. Extant timetrees are consistent with a myriad of
601 diversification histories. Nature. 580:502–505.
602 Manceau M., Marin J., Morlon H., Lambert A. 2020. Model-based inference of
603 punctuated molecular evolution. Mol. Biol. Evol.:msaa144.
604 Mayr E. 1954. Change of genetic environment and evolution. Evol. as a
605 Process.:157–180.
606 Mindell D.P., Sites J.W., Graur D. 1989. Speciational Evolution: a Phylogenetic
607 Test With Allozymes in Sceloporus (Reptilia). Cladistics. 5:49–61.
608 Mindell D.P., Sites J.W., Graur D. 1990. Mode of allozyme evolution: Increased
609 genetic distance associated with speciation events. J. Evol. Biol. 3:125–131.
610 Nabholz B., Glémin S., Galtier N. 2008. Strong variations of mitochondrial
611 mutation rate across mammals - The longevity hypothesis. Mol. Biol. Evol.
612 25:120–130.
613 Nee S., May R.M., Harvey P.H., Trans P., Lond R.S. 1994. The reconstructed

evolutionary process. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 344:305–11.

Pagel M., Venditti C., Meade A. 2006. Large punctuational contribution of speciation to evolutionary divergence at the molecular level. Science. 314:119–121.

Pybus O., Harvey P. 2000. Testing macro–evolutionary models using incomplete molecular phylogenies. Proc. R. Soc. B Biol. Sci. 267:2267–72.

Rabosky D.L., Donnellan S.C., Talaba A.L., Lovette I.J. 2007. Exceptional among-lineage variation in diversification rates during the radiation of Australia's most diverse vertebrate clade. Proc. R. Soc. B Biol. Sci. 274:2915–2923.

Ricklefs R.E. 2006. Global variation in the diversification rate of passerine birds. Ecology. 87:2468–2478.

Russel P.M., Brewer B.J., Klaere S., Bouckaert R.R. 2019. Model Selection and Parameter Inference in Phylogenetics Using Nested Sampling. Syst. Biol. 68:219–233.

Saclier N., François C.M., Konecny-Dupre L., Lartillot N., Guéguen L., Duret L., Malard F., Douady C.J., Lefébure T. 2018. Life history traits impact the nuclear rate of substitution but not the mitochondrial rate in isopods. Mol. Biol. Evol. 35:2900–2912.

Schliep K.P. 2011. phangorn: phylogenetic analysis in R. 27:592–593.

Simpson G.G. 1945. SECTION OF BIOLOGY: Tempo and Mode in Evolution. Trans. N. Y. Acad. Sci. 8:45–60.

Sipos B., Massingham T., Jordan G.E., Goldman N. 2011. PhyloSim - Monte Carlo simulation of sequence evolution in the R statistical computing environment. .

Spielman S.J., Wilke C.O. 2015. Pyvolve: A Flexible Python Module for

639 Simulating Sequences along Phylogenies. :1–7.
640 Stadler T. 2011. Simulating trees with a fixed number of extant species. Syst. Biol.
641 60:676–684.
642 Sung W., Ackerman M.S., Dillon M.M., Platt T.G., Fuqua C., Cooper V.S., Lynch M.
643 2016. Evolution of the insertion-deletion mutation rate across the tree of
644 life. G3 Genes, Genomes, Genet. 6:2583–2591.
645 van Valen L.M. 1985. Why and how do mammals evolve unusually rapidly. Evol.
646 Theory. 7:127–132.
647 Venditti C., Meade A., Pagel M. 2006. Detecting the node-density artifact in
648 phylogeny reconstruction. Syst. Biol. 55:637–643.
649 Venditti C., Pagel M. 2010. Speciation as an active force in promoting genetic
650 evolution. Trends Ecol. Evol. 25:14–20.
651 Webster A.J., Payne R.J.H., Pagel M. 2003. Molecular phylogenies link rates of
652 evolution and speciation. Science. 301:478.
653 Zhang R., Drummond A. 2020. Improving the performance of Bayesian
654 phylogenetic inference under relaxed clock models. BMC Evol. Biol. 20:1–28.
655