Understanding cultural evolution in hummingbird leks through the fossilized birth-death process

Marcelo Araya-Salas (marcelo.araya@ucr.ac.cr)[1,2] *

Beatriz Willink (beatriz.willink@ucr.ac.cr)[3,4]

Alejandro Rico-Guevara (colibri@uw.edu)[5]

2024-10-02

```
## [1] "475 duplicate(s) references found in combined_bibs.bib"
```

[1] Centro de Investigación en Neurociencias, Universidad de Costa Rica

[2] Lab of Ornithology, Cornell University

[3] School of Biology, University of Costa Rica

[4] Department of Biology, Stockholm University

[5] Department of Biology, University of Washington

*To whom correspondence should be addressed

Keywords:

Contents

Abstract

Introduction

The idea that socially transmitted behaviours change and diversify over time in a manner analogous to organic evolution and phylogenetic diversification can be traced back to Darwin, who noted that "the formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel" (Darwin 1871). The term culture can be used broadly to refer to socially transmitted information that influences behavioural patterns within animal groups (Laland and Hoppitt 2003). While human language, beliefs, norms and material artefacts are well-known cultural domains, many forms of culture exist among diverse animals, such as vocal dialects (Catchpole and Slater 2003; Aplin 2019), navigation routes (Laland and Williams 1997; Jesmer et al. 2018), and tool use traditions (Whiten et al. 2005; Luncz and Boesch 2014). After the Modern Synthesis, the formal modelling of cultural change as evolution, whether in humans or non-human animals, became possible by drawing analogies between cultural and population genetic processes (Cavalli-Sforza and Feldman 1981; Boyd and Richerson 1985). For instance, imperfect imitation could be compared to genetic mutation (Kempe et al. 2012), biased transmission to natural selection (Williams et al. 2013) and random fluctuations in the frequency of traditions to genetic drift (Bentley et al. 2004). However, unlike genetic evolution, in which the most fundamental units of transmission (nucleotides) are essentially universal, cultural evolution implies disparate units of transmission across taxa and in different social contexts (e.g. tools vs. songs). To properly address the long-standing question of whether cultural change over time is truly akin to evolution, we require means to systematically assess the power of evolutionary methods, across the great variety of cultural forms that have emerged in the history of animals (Mesoudi 2017).

Some behaviours can be described as sequences of ethological units. For example, visual displays can be encoded as a string of stereotyped motor patterns (Ligon et al. 2018; Araya-Salas et al. 2019) and bird and whale songs are typically structured as sequences of repeated and hierarchically nested sounds (Payne and McVay 1971; Rivera-Cáceres et al. 2016; Kershenbaum et al. 2016; Garland et al. 2017). Encoding behaviour as a sequence can facilitate the adoption of phylogenetic approaches that take molecular data as their main input. Such approaches have been developed within a strong theoretical framework that continues to grow and increasingly accommodates biological realism (Yang and Rannala 2012). Substitution models applied to molecular sequence evolution are routinely combined with clock models and tree priors, such as the birth-death process, to understand the temporal dynamics of lineage diversification and turnover (Morlon 2014; Bromham et al. 2018) Analogously, clock models, tree priors and substitution models may be used to elucidate the temporal dynamics of cultural diversification, when culture can be adequately modeled as behavioural sequences composed of discrete units. However, most phylogenetic models assume that once

diverged, lineages evolve independently of one another, branching in a tree-like pattern. As in organisms with extensive hybridization and horizontal gene transfer (Philippe and Douady 2003; Bapteste et al. 2013), this assumption may be often violated in cultural evolution, because learning from same-cohort individuals should increase the potential for reticulate evolution (Gray et al. 2007). Empirical knowledge on the robustness of classic phylogenetic models to horizontal transmission of culture (e.g. Collard et al. 2006), and more generally the absolute fit of such models models to cultural data, is thus crucial in addressing the utility of phylogenetic approaches to study the cultural diversification of socially learnt behaviours.

Cultural evolution poses further challenges to the application of phylogenetic models. Most implementations of phylogenetic models on molecular data start with a sequence alignment. Alignments represent assumptions of homology between characters in matched positions along a sequence, but are typically treated as observations for phylogenetic inference (Lutzoni et al. 2000; Redelings and Suchard 2005; Lunter et al. 2005). Numerous methods of sequence alignment have therefore been developed to capture the main features of molecular evolution, and in some cases to explicitly model substitution events (Yang and Rannala 2012; Chatzou et al. 2016). Nonetheless, the accurate reconstruction of homology in sequence alignments is a pervasive challenge in molecular phylogenetics (Warnow 2021), that is only exacerbated when borrowing phylogenetic tools for the study of behavioural sequences (Caetano and Beaulieu 2020). We clearly have a better understanding of the basic rules that govern the rates of different nucleotide substitutions than we do for changes in the dance moves of a courtship display or changes in the sequence of sounds of a mating call. A crucial question for the nascent field of cultural phylogenetics (sensu Mesoudi 2017) is therefore whether alignment algorithms developed for molecular data can be suitably modified to represent the processes behind cultural change.

Despite these challenges, culture also poses unmatched opportunities for the application of phylogenetic inference. Culture can change very rapidly in comparison to molecular evolution (Perreault 2012), allowing researchers to document lineage diversification events as they occur and during the span of one or a few academic lifetimes. Cultural phylogenetics can therefore capitalize on a relatively rich historical record, that markedly contrasts the sparse fossil record of many organismal groups (Kidwell and Holland 2002). Recently developed phylogenetic methods have shown that sampling ancestors of extant taxa and explicitly incorporating these data in the diversification process allow for more accurate estimation of divergence times (Gavryushkina et al. 2014, 2017; Zhang et al. 2016). This is accomplished by the fossilized birth-death process (FBDP) (Heath et al. 2014; Gavryushkina et al. 2014), a model that jointly describes the probabilities of lineage splitting, extinction and fossilization that give rise to the sampled taxa, whether extant or fossil. Of course, the fossilization rate estimated in the FBDP may represent actual fossilization events, but can also be used to describe serially sampled viral strains (Stadler and Yang 2013; Gavryushkina et al. 2014), or,

as in this case, historical records of behavioural patterns that are socially learnt and transmitted (Rama 2018; Ritchie and Ho 2019; Zhang et al. 2020). Thus, when culture evolves rapidly and learnt behaviours are sampled serially, a vast record of ancestral lineages can bolster inferences of cultural diversification dynamics through the FBDP.

Cultural phylogenetics research that builds on Bayesian estimation of origination, extinction and preservation rates is recently growing, but remains restricted to specific domains of human culture (Gjesfjeld et al. 2016, 2020; Rama 2018; Ritchie and Ho 2019; Sagart et al. 2019; Zhang et al. 2020). Studies applying the FBDP in particular have been focused on elucidating the history of diverse human language families (Rama 2018; Sagart et al. 2019; Zhang et al. 2020). Thus, a great untapped potential remains for investigating cultural diversification through the FBDP in non-human animals. Socially learnt bird songs, such as in the long-billed hermit (*Phaethornis longirostris*; Fig. 1a) are obvious candidates to examine the suitability of the FBDP as a phylogenetic model of cultural diversification. Evidence of social learning in this species (**TenCate2021a?**), includes micro-geographic song variation decoupled from genetic structure (Araya-Salas et al. 2019) and adult replacement of crystallized songs (Araya-Salas and Wright 2013) . The long-billed hermit song can be represented as a sequence of discrete sounds fused together into an unbroken signal (see 'Methods'). Indeed, the most salient differences among song types reside in the composition and sequential order of their sounds (Araya-Salas and Wright 2013). Males sing a single song-type repertoire, which enables comparisons of individual songs as homologous traits (as opposed to multiple song-type repertoires). Courtship occurs within leks of 5-20 highly vocal males (**stiles-wolf1979?**), which facilitates longitudinal monitoring of all song types within a lek. Moreover, song types can be shared by sub-groups of males within leks, with no evidence of song type sharing across leks (Araya-Salas et al. 2019), suggesting that leks operate as relatively isolated cultural systems. Limited inter-lek migration also provides an opportunity to gain unique insights on the repeatability of cultural evolution when song changes are monitored in multiple leks.

Here, we used the FBDP to model cultural diversification in five leks of Long-billed hermits, using historical song surveys spanning up to five decades (Fig. 1c-d). We then investigated model reliability and absolute fit of phylogenetic models, using posterior predictive simulation and comparing features of empirical song sequences to sequences generated by models under the FBDP. We further asked how biologically informed assumptions during sequence alignment impact model reliability and estimates of diversification dynamics. Finally, we explored how the use and completeness of historical records (analogous to fossil records) affect model reliability and the fit of alternative clock models to long-billed hermit song data. Our results shed light into the adequacy of historically-informed phylogenetic inference for reconstructing cultural change in non-human animal systems.

—>

Figure 1. Socially transmitted songs in the lekking long-billed hermit. **a)** A male Long-billed hermit. **b)** Spectrograms of three songs from different males in the SUR lek, sampled in 2019. The *colour* arrow head shows a pure tone and the *another colour* arrow head shows a vibratory sound. **c)** Locations of the study leks in the Caribbean lowlands of Costa Rica. **d)** Sampling of historical song records in the study leks.

Methods

Song collection and coding

Song recordingos come from five leks and four sites in Eastern Costa Rica (Fig. 1c). All sites were surveyed between 2008 and 2019 [Araya-Salas REF?], and additional historical recordings going back decades up to five decades were obtained for the leks in La Selva and Hitoy Cerere (Fig. 1d). Recordings were gathered with different equipment at different points in time (i.e. shotgun or parabolic microphones, analog or digital recorders). Nonetheless, the spectrographic structure of the signals (used for determining signal composition, see below) is not affected by the recording equipment in a significant manner. The most noticeable effect of differences in recording equipment can be a slight time distortion (expansion or contraction) when using analogous recordings [REF?]. However, the approach used for coding song structure (explained below) as sequences is not affected by song duration (i.e. a time-expanded song would produce exactly the same song

sequence as its original form).

Long-billed hermit songs are composed of two basic type of sounds: pure tones and trills (Fig. 1b, S1). Pure tones can vary in the degree of modulation (i.e. changes in frequency through time) while trills vary in the number of oscillations per unit of time (i.e. rate). We subdivided these two basic sound types into six categories (Fig. S1): slow trill, medium-paced trill, fast trill, downward pure tone, upward pure tone and flat pure tone. Song were split into 20 equal-length segments and each segment was assigned to one of these six categories (Fig. S1).

**We need to explain two things here: 1. How do we justify splitting a song in 20 equal-length intervals 2. How did we repeatably classified these inervals

I suggest a quick description here and a lengthy description in the supporting material including a schematic figure based on this "Fig. 1"**

Sequence alignment

Alignment of behavioural sequences is complicated by the challenge of establishing homology between ethological segments or units (Caetano and Beaulieu 2020). Here, we implemented and compared three alignment strategies based on two methods originally developed for multiple sequence alignment (MSA) of molecular data. In alignments of nucleotide and protein sequences, gaps represent insertion or deletion mutations, so that characters at gapped sites lack homology across the dataset. Commonly used MSA methods differ in their treatment of insertion and deletion events in ways that can impact homology inferences in cultural as well as in molecular characters (Löytynoja 2012). MAFFT (Katoh et al. 2002; Katoh and Standley 2013) uses a progressive alignment algorithm with a default gap-opening penalty (1.53) and no gap extension penalty by default, in versions > 6.626. The L-INS-i method follows the progressive alignment by iterative refinement, based on consistency and weighted sum-of-pairs scores. In MAFFT versions $>$ 7.371 user-defined alphabets and scoring matrices can be implemented in addition of nucleotide and amino acid alternatives. MAFFT is therefore a flexible program to align behavioural sequences in which changes analogous to multi-site insertions and deletions have occurred, and which are composed by a variable number of character states. In our first alignment strategy, which we hereafter refer to as 'MAFFT-agnostic', we used the MAFFT L-INS-i method with default gap penalties and a customized scoring matrix in which all transitions between alternative character states were equally likely.

Our second alignment strategy also used the MAFFT L-INS-i method and default gap penalties, but we made the assumption that when long-billed hermits modify pre-existing songs they are more likely to replace a trill

by a different type of trill and a tone by a different type of tone than to change from vibratory to pure tones or *vice versa*. This seems more biologically meaningful as different biomechanics are involved in producing trills and pure tones (**Elemans2007?**). We implemented this assumption by enforcing a higher cost of mismatches between sound categories than within either trills or pure tones. To determine an appropriate difference in mismatch scores, we capitalized on a previously documented pattern of higher song similarity within lek than between leks (Araya-Salas et al. 2019), and the observation of no song type sharing among four leks in close proximity during a ten year period (M.A.S. pers. obs.), suggesting little cultural transmission across leks. . We would thus expect longer alignments in data sets composed of song sequences from different leks than in data sets composed of sequences from the same lek, as these sequences have a more recent common ancestor. Following this logic, we selected mismatch scores for substitutions within and between sound categories (trill vs. pure tone) that maximize the alignment length for pools of sequences from different leks, relative to the alignment length sequences originating from the same lek. Mismatch score optimization was based on a data set of 184 song sequences from 12 leks, including the focal five leks of the present study. We hereafter refer to this alignment strategy as 'MAFFT-optimal'.

For our third alignment strategy, we used the phylogenetically informed alignment program PRANK (Löytynoja and Goldman 2005, 2008). PRANK also uses a progressive algorithm but handles the placement of insertions and deletions differently, by using outgroup information in the subsequent alignment step. PRANK thus uses the sequence phylogeny to differentiate insertions from deletions, and thereby avoids site overmatching by penalizing insertions in a single stage of the alignment (Löytynoja and Goldman 2005). Unlike MAFFT, PRANK is an evolutionary aware program in that insertions, deletions and substitutions are modelled explicitly on a phylogenetic tree. However, PRANK does not currently support customized alphabets and substitution-rate matrices. To use PRANK we assumed that flat tones can be treated as ambiguous between upward and downward tones, and medium-paced trills can similarly be treated and ambiguous between fast and slow trills. We therefore used IUPAC ambiguity code for DNA nucleotides to rename song segments, with tones as purines and trills as pyrimidines. As per PRANK's defaults we used a TN93 nucleotide substitution model with empirical base frequencies and transition/transversion rate ratio (κ) = 2. Therefore, as in the 'MAFFT-optimal' alignment, in the 'PRANK-TN93' alignment we explicitly assumed a higher transition rate within vibratory and pure sound categories than between them. For this alignment we used the default gap-opening rate and extension probabilities (0.025 and 0.75 respectively) and we omitted the -F option that fixes inferred insertions but increases sensitivity to guide-tree accuracy.

Phylogenetic analysis

All phylogenetic analyses were conducted in RevBayes v. 1.0.12 and v. 1.1.0, a computation environment that uses probabilistic graphical models for Bayesian inferences in phylogenetics and evolution (Höhna et al. 2016). Our phylogenetic model was a fossilized birth-death process (FBDP) which describes the joint prior distribution of the tree topology, divergence times and lineage sampling times before the present (Heath et al. 2014). In the FBDP, extant taxa and lineages sampled before the present are part of the same macroevolutionary process. For many applications of the FBDP, extinct or ancestral taxa can only be sampled through fossils. However, in the case of fast-evolving songs that are culturally transmitted among individuals, historical records of songs are equivalent to fossil data. Historical records contain the character sequences of songs that existed in the past and may be ancestral to extant songs or may have gone extinct. As in the FBDP with fossil data, the probability that a historical song is an ancestor of extant songs depends on the rates of lineage turnover and the rate of recovery of historical records (hereafter recovery rate). The recovery rate is the rate at which ancestral songs are sampled from the lineage diversification process and it is a random variable drawn from a prior distribution, such as the birth and death rates of a traditional birth-death model. Sampling ancestors as part of the same evolutionary process as we have done here improves estimation of diversification and clock rates (Gavryushkina et al. 2014), and sampling character data from ancestors further improves estimates of divergence times in simulated data (Luo et al. 2020).

Our dataset on historical records of songs in hummingbird leks has three advantages in comparison to most fossil datasets used in phylogenetic analyses. First, there is no stratigraphic uncertainty. We can be certain that historical songs occurred in the year when they were recorded. Second, there are no partial fossils. Songs recorded in the past are just as complete as the most recent ones, creating no additional ambiguity in character states of historical songs. Third, the historical record is relatively rich. In all leks, there are multiple years sampled consecutively, and in two leks (SUR and CCE), historical records go back to 1969 (Fig. 1d). Because leks are small and long-billed hermits are actively singing their single song type throughout the breeding season, we can assume detection is nearly perfect. We therefore do not need to account for missing taxa in any of our phylogenetic analyses.

A possible complication in our analysis is that sampling years are interspaced by long gaps without recordings in the three leks with deeper historical surveys (HC1, CCE and SUR). However, the temporal distribution of these ancestral samples is not unlike that of fossils, which are typically aggregated in discrete strata of exposed rocks (Holland 2016). The FBDP is robust to some forms of bias in fossil sampling, including non-continuous recovery (Heath et al. 2014). Nonetheless, to better understand the effects of deep, yet discontinuous historical sampling we conducted all analyses for these leks both with the complete dataset,

including long gaps without lineage sampling, and with the more recent and continuously sampled dataset. We present both sets of results for comparison. Finally, to investigate the general impact of sampling historical records on phylogenetic inference of song evolution, we conducted an additional set of analyses, including only songs observed in the last year of sampling (i.e. analogous to sampling only extant taxa). For these analyses without historical records, we used the three leks (BR1, SUR and TR1) that had 3 or more distinct songs in their last year of sampling.

Another potential issue arises in the years with highly frequent sampling, in which identical songs could be sampled at multiple time points. This is uncommon for fossil data, as it would entail the discovery of fossils with the same character state combination in multiple horizons. Here, we focus on the results of analyses in which all historical occurrences are considered in the evolutionary process, including identical songs sampled in consecutive years. However, we also conducted all analyses accounting only once for each unique song, at its earliest occurrence. The results of these analyses with only the earliest occurrence of songs are presented in the Supporting Material.

Phylogenetic analyses were conducted with all three alignment strategies (MAFT-agnostic, MAFT-optimal and PRANK-TN93) for each lek. We used an exponential prior with rate parameter = 10 for the speciation, extinction and historical sampling rates, and a broad uniform prior, bounded between 1000 and 0 years, on the root age of all leks. Song sequences were assumed to evolve under a generalised time-reversible (GTR) model with exchangeability rates and stationary frequencies drawn from a flat Dirichlet prior. Site-rate heterogeneity was modelled with a discretised gamma distribution with four rate categories and with equal shape and scale parameters, in turn drawn from an exponential prior with rate = 10.

We tested both global and relaxed clocks for song evolution. Branch rates under the global clock were drawn from an exponential prior with rate = 10. Branch rates under the relaxed clock were uncorrelated and drawn from an exponential prior, with mean in turn drawn from an exponential hyperprior with rate = 10. We compared clock models using marginal likelihood approximation via the stepping stone algorithm (Xie et al. 2010). Clock-model comparisons were conducted for each lek (BR1, CCE, HC1, SUR, TR1), alignment (MAFFT-agnostic, MAFFT-optimal, PRANK-TN93), historical dataset (oldest records included, recent records only, no fossils) and use of historical records per song (using all, using earliest). For diversification dynamics, tree comparisons and tests of model reliability (see below), we present results under the preferred clock model here and for the alternative model in the Supporting Material.

We conducted two independent MCMC runs for each analyses, with 150,000 generations and an additional 50,000 of burn-in for leks with fewer song types (BR1, HC1, TR1) and 300,000 generations of posterior

sampling and 100,000 generations of pre-burnin for the two largest leks (CCE and SUR). In all cases parameter tuning was conducted every 200 generations. To improve mixing, we used the Metropolis-Coupled MCMC sampler with three heated chains and default swapping parameters. To avoid autocorrelation in the posterior we saved samples every 100th generation. We assessed MCMC performance using the package 'coda' (Plummer et al. 2006) in R v. 4.0.4 (R Core Team 2021). We checked for convergence between independent runs visually and using the Gelma-Rubin potential scale-reduction factor (psrf). We assumed convergence if psrf < 1.05 for all variables. We also inspected autocorrelations between draws (targeted below 0.1) and effective sample sizes (targeted above 200) for all model variables. We summarise MCMC diagnostics in the Supporting Material.

Model reliability

We used predictive data simulations to test for absolute model fit, also implemented in RevBayes (Höhna et al. 2018). During parameter inference, a Stochastic-Variable-Monitor stored the stochastic variable values for each posterior sample. Then, these values were used to simulate new datasets based on the inference model. We specified a thinning of 2 iterations for the stochastic variable trace of each independent chain, thus simulating 3 000 datasets for the 'large' models (CCE, SUR) and 1 500 datasets for the 'small' models (BR1, HC1, TR1).

We present data-based test statistics comparing simulated to empirical datasets, as tests of absolute model fit. We calculated 10 such statistics: 1) the number of invariant sites in the alignment, 2) the number of segregating sites in the alignment, 3) the maximum length of invariant blocks, 4) the maximum length of variable blocks, 5) the number of invariant blocks, 6) the maximum pairwise difference between two sequences in an alignment, 7) the minimum pairwise difference between two sequences in an alignment, and three measurements of genetic diversity: 8) Watterson's θ, an estimate of "population mutation rate" (Watterson 1975), 9) Tajima's D, a measurement of whether a population evolves neutrally (Tajima 1989), 10) π, the average number of pairwise differences in the alignment, also used to calculate Tajima's D. For more details about these statics and how they are calculated see Höhna et al.(2018).

For each test, we report a posterior predictive effect size (PPES) and a two-tailed posterior predictive p-value (Höhna et al. 2018). The PPES of each statistic corresponds to the difference between the median of the posterior distribution of simulated data sets and the empirical value, normalized by the SD of the posterior distribution (Höhna et al. 2018). The two-tailed posterior predictive p-value is calculated by first obtaining a lower and upper tail p-value and multiplying the smaller of the one-tailed p-values by two. The lower one-tailed p-value is the proportion of simulated data sets in which the value for the test statistic is less than or equal to the observed value. The upper one-tailed test is the proportion of simulated data sets in which

the value for the test statistic is greater than or equal to the observed value. Especially with small data sets, it is possible that test statistics in mutliple simulated data sets are equal to test statistics in the empirical data. In these cases the smaller of the two one-tailed p-values could be greater than 0.5. In these cases we report a value of 1 as the posterior predictive p-value.

Treespace and parameter sensitivity

We explored tree topology congruence of different models by comparing topological distances between high posterior probability trees. Topologies were compared with the Robinson-Foulds distance (**Robinson?**) with the R package phangorn (**Schliep2011?**). Only tree tips shared by all trees were included in the analysis. Topological distances were projected in a bidimensional space using Classic Multidimentional Scalling in order to quantify topological space. We estimated the overall spread of the topological space (i.e. space size) for different models as a metric of within-model topological congruence . We also calculated between-model topological congruence as the overlap of the topological space of a model to the spaces from other models. Space overlap was estimated as the proportion of the joint area of two spaces that was shared. Topological congruence descriptors were calculated using the R package PhenotypeSpace (**Araya-Salas2022?**). Within-model congruence was mean-centered by lek to allow comparisons across leks. The effect of different model specifications on these two topological space descriptors was evaluated using Bayesian lineal regression models with each descriptor as the response variable and model alignment strategy, use of historical data, historical record completeness, and clock model as predictors. Regression models were run in Stan (**Carpenter2017?**) through the R platform (**Team2021?**) using the package brms (Bürkner 2017). We present effect sizes as median posterior estimates and 95% credibility intervals (CI) as the highest posterior density interval. Parameters in which credible intervals did not include zero were regarded as having an effect on the response variable. Models were run on three chains for 2500 iterations, following a warm-up of 2500 iterations. Effective sample size was kept above 3000 for all parameters. Performance was checked visually by plotting the trace and distribution of posterior estimates for all chains. We also plotted the autocorrelation of successive sampled values to evaluate independence of posterior samples. Potential scale reduction factor was used to assess model convergence and kept below 1.05 for all parameter estimates.

We also asked if inferences of diversification dynamics and song evolution were influenced by the use of different alignment strategies. To do this we compared the posterior distributions of parameter estimates between the MAFFT-agnostic, MAFFT-optimal and PRANK-TN93 alignment strategies. For diversification dynamics we compared speciation and extinction rates, as well as the age of the MRCA of all songs (including extinct lineages) and the age of the MRCA of only extant songs (those present in the last year of sampling). For song

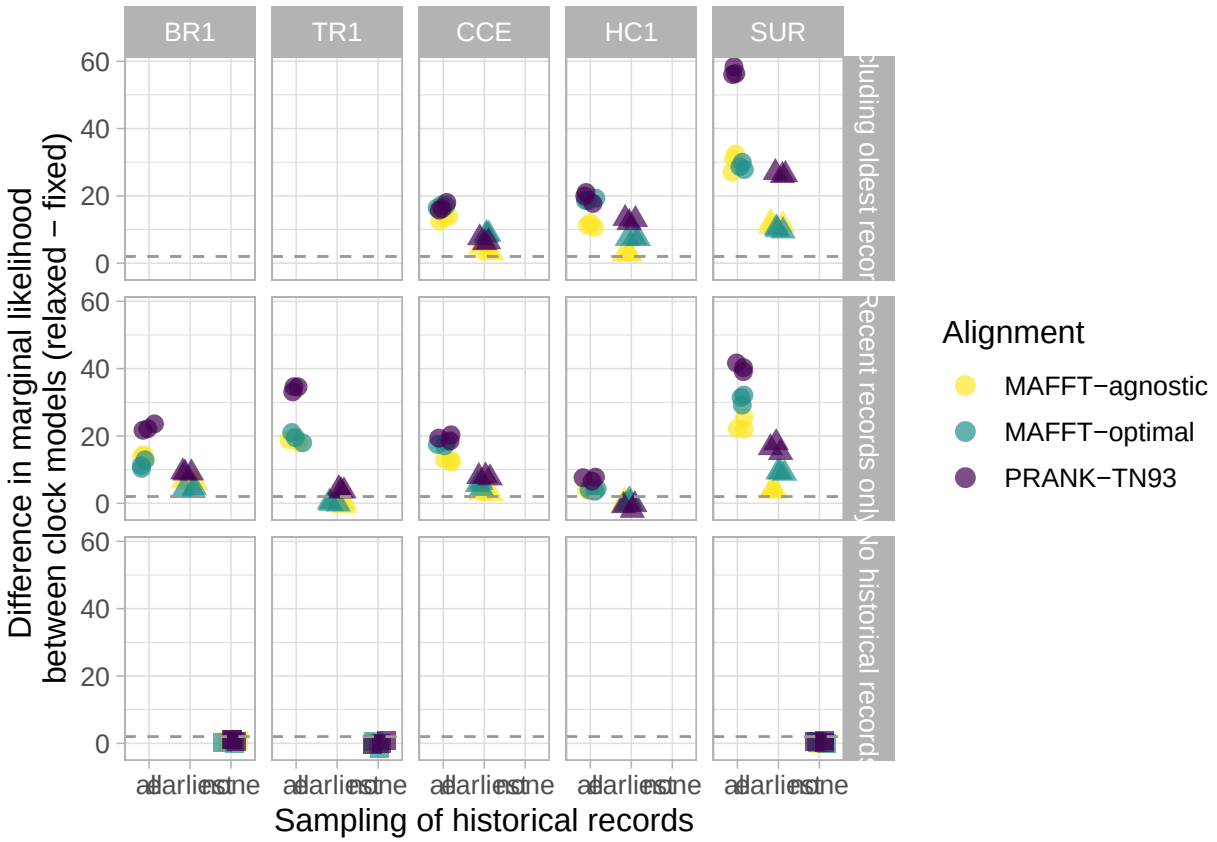
evolution, we focused on substitutions rates between broad sound categories (trill to pure tone and vice versa), substitution rates within sound categories (e.g. between fast and slow trills), and stationary frequencies of sound categories. We considered parameter estimates to differ between a "query" and "target" alignment strategies if more than 5% of the posterior distribution of the parameter under the "query" alignment fell outside the95% highest posterior density (HPD) interval of the parameter estimate under the "target" alignment. We similarly explored sensitivity of diversification parameters to fossil use and clock models.

Results

Model reliabilty

Clock model selection

Support for a relaxed clock model of song evolution (in which different son lineages evolve at different rates) depended on the historical record and sampling strategy. When historical data was excluded, there was no increase in ML by relaxing the clock model. However, the use of historical songs akin to fossils resulted in a higher fit of the relaxed model, particularly when all historical records, including identical song sequences sampled in consecutive years are incorporate in the macroevolutionary process. While this trend was present in most leks and data sets, it tended to be stronger under the PRANK-TN93 alignment, especially in the historically largest lek (SUR)

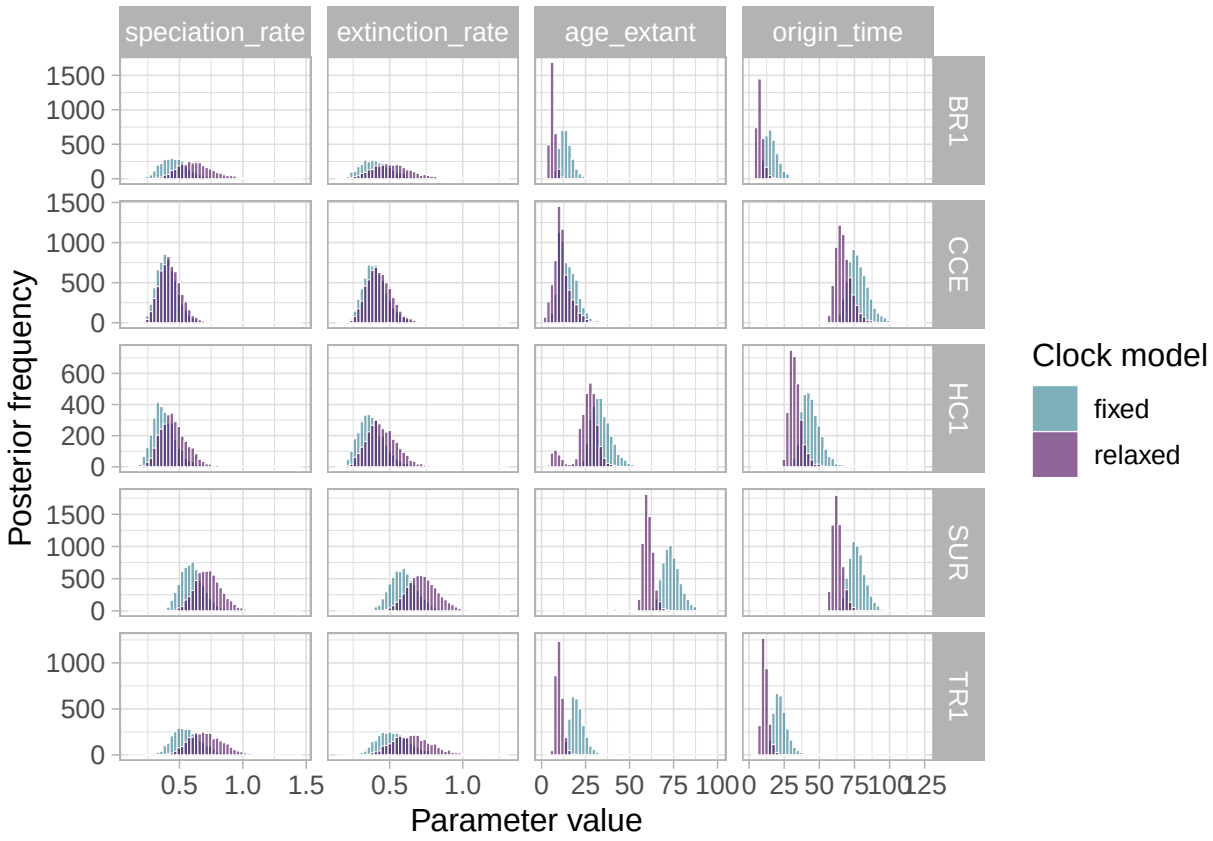


Treespace congruence

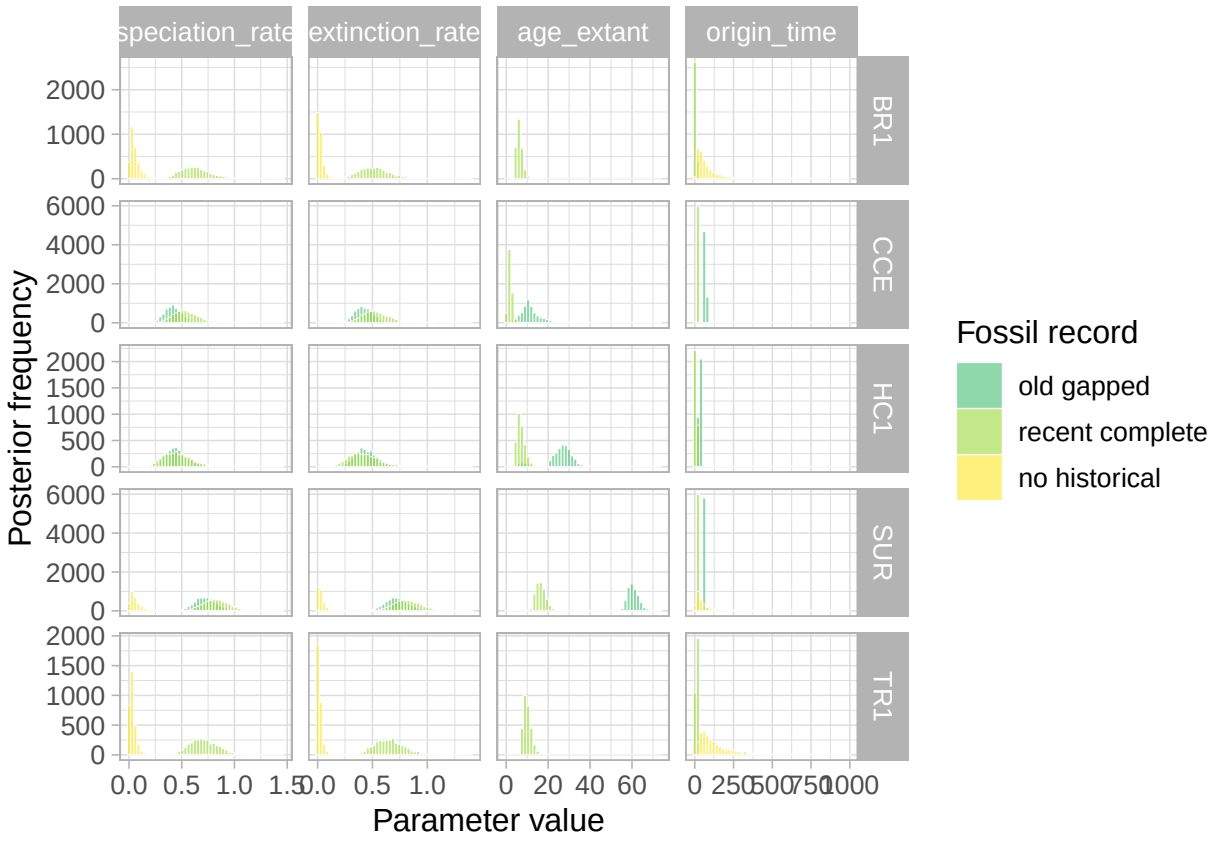
The congruence of topologies within a model was only affected by the use of historical records: the spread of the topological space for tree tips shared across all models increased when including historical data (effect size: 0.02, 95% CI= 0.01 _ 0.02). Topological congruence between models was lower when using the PRANK-TN93 alignment strategy compared to both MAFFT-agnostic (effect size: -0.09, 95% CI= -0.13 _ -0.06) and MAFFT-optimal strategies (effect size: -0.06, 95% CI= -0.10 _ -0.03). The use of historical records also generated decreased topological congruence to other models (effect size: -0.11, 95% CI= -0.14 _ -0.08).

Diversification dynamics

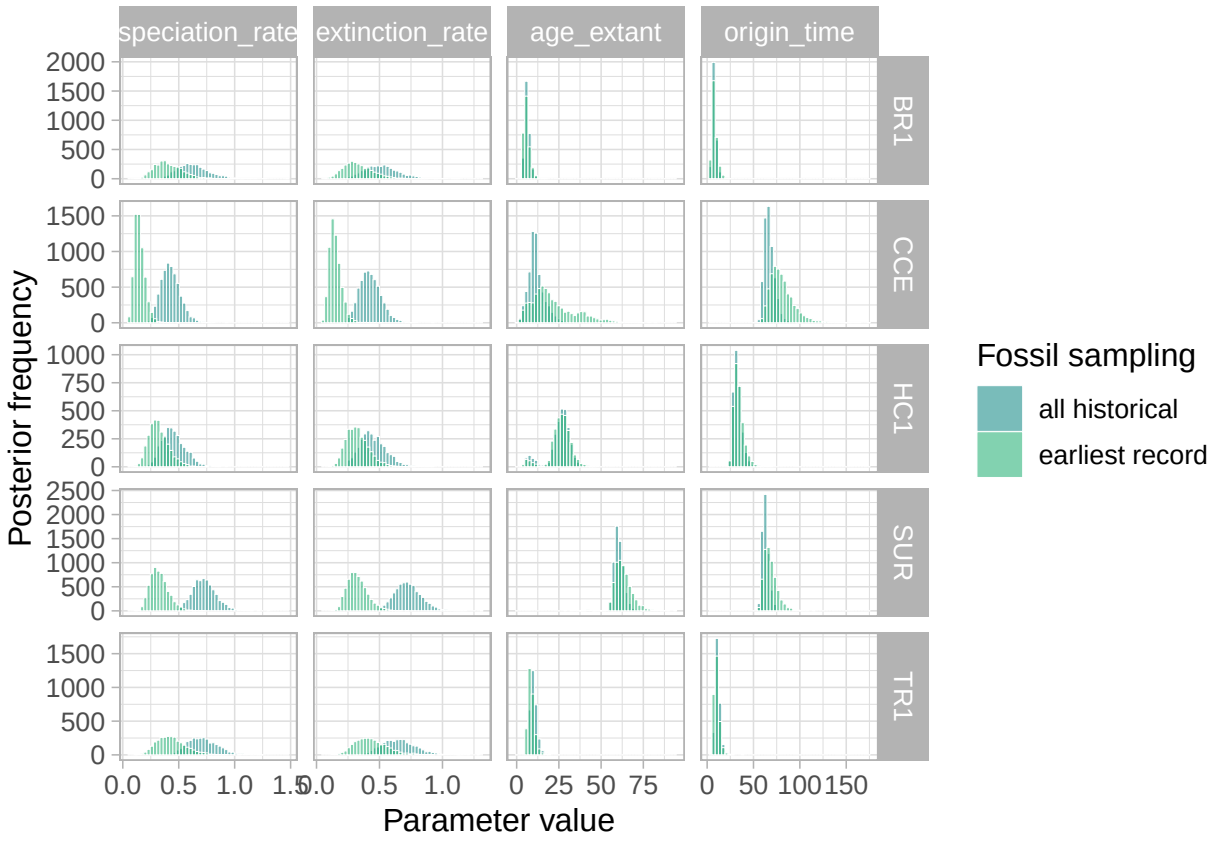
Diversification rates were not super sensitive to alignment strategy, with exception of SUR, where PRANK infers slower turnover. Age of extant taxa is VERY sensitive to alignment for leks with large gaps in their historical record (CCE, HC1, SUR). In SUR, only PRANK results imply that more than a single lineage from earliest sampling years has survived to the present. In HC1 both PRANK and MAFFT-optimal make this inference and in CCE, all alignments result in a recent origin of extant songs. Origin times of all songs are not so sensitive to alignment strategy, but PRANK tends to infer older ages.



Diversification rates were also very sensitive to historical information. In the three leks in which analyses without fossils could be conducted (BR1, TR1, SUR), fossil-free inference resulted in much slower diversification rates and markedly uncertain origin times in comparison to analyses including historical songs. For the three leks in which a long but gapped historical record could be contrasted with a shorter but complete one (CCE, HC1 and SUR) including more ancient records resulted in older age estimates for the MRCA of both extant and extant + historical songs. Here we show this for PRANK, but the same pattern holds for alternative alignments (Supporting Material).



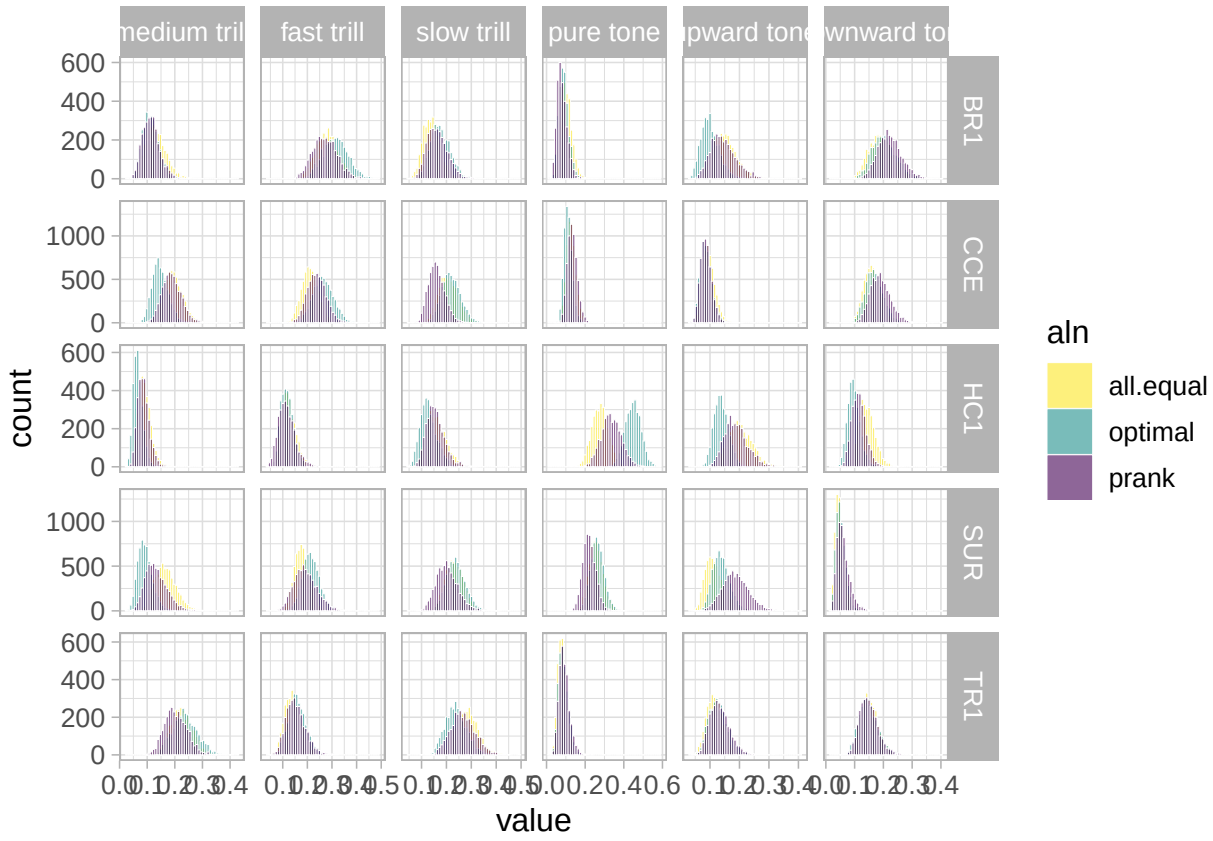
Diversification inference was also sensitive to how consecutive historical records are incorporated into the analysis. Including all consecutive samplings of identical songs resulted in higher turnover (higher speciation and extinction). Generally, age estimates were not so sensitive, with the exception of CCE and SUR (leks with longer historical sampling), in which considering only the earliest appearance of a historical song caused greater uncertainty and older estimates in the origin time of all songs (extant and historical). Here we show this for PRANK, but the same pattern holds for alternative alignments (Supporting Material).



374

375 **Song evolution**

376 We looked at how the parameters of substitution models were influenced by alignment strategies. Different
377 alignment strategies create different assumptions about sequence homology. Here leks are VERY idiosyncratic.
378 The stationary frequencies of sound types vary across leks and they can be highly sensitive to the alignment
379 strategy in some leks but not others. For example, pure tones are more frequent in HC1 than in other leks,
380 but the estimated frequency varies markedly among alignment strategies, whereas in TR1 different alignments
381 are congruent. Similarly, the PRANK alignment results in a higher frequency of medium trills and a lower
382 frequency of fast and slow trills in SUR and CCE, compared to MAFFT-optimal. However in CCE and TR1
383 PRANK alignment results in a higher frequency of slow trills.

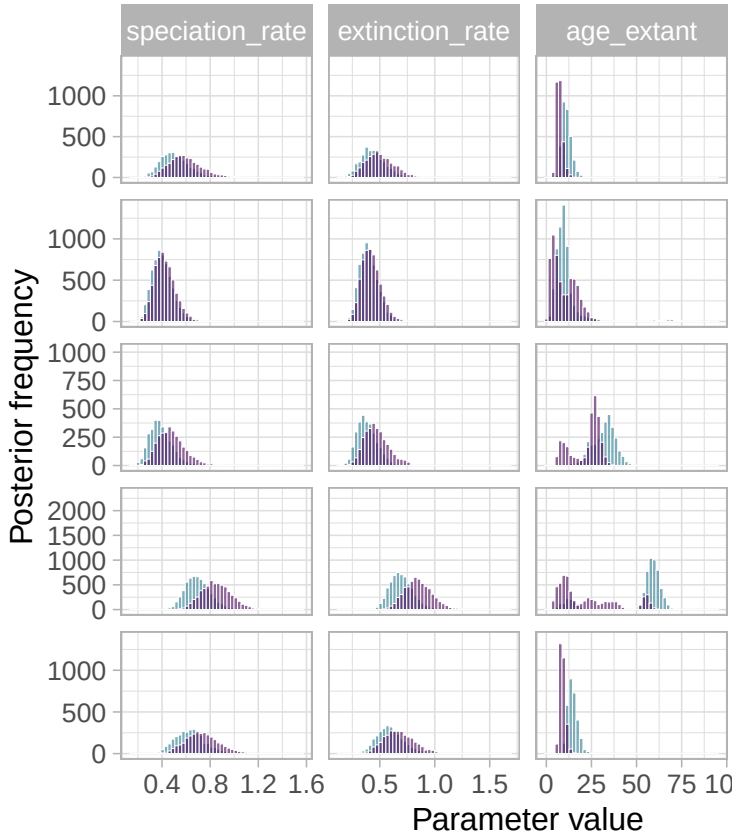


384

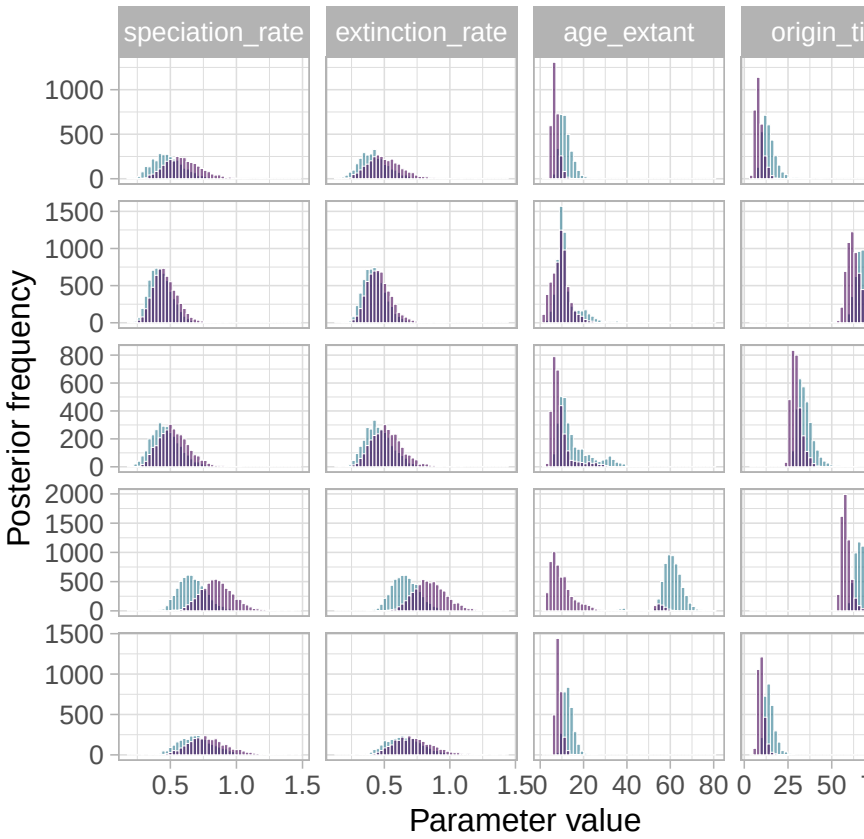
385 A similar scenario arises in substitution rate estimates, where the sensitivity of rate estimates and the effects
386 of particular alignment strategies vary vastly across leks, with HC1 being particularly prone to incongruence
387 between alignment strategies (Supporting Material?).

Discusion

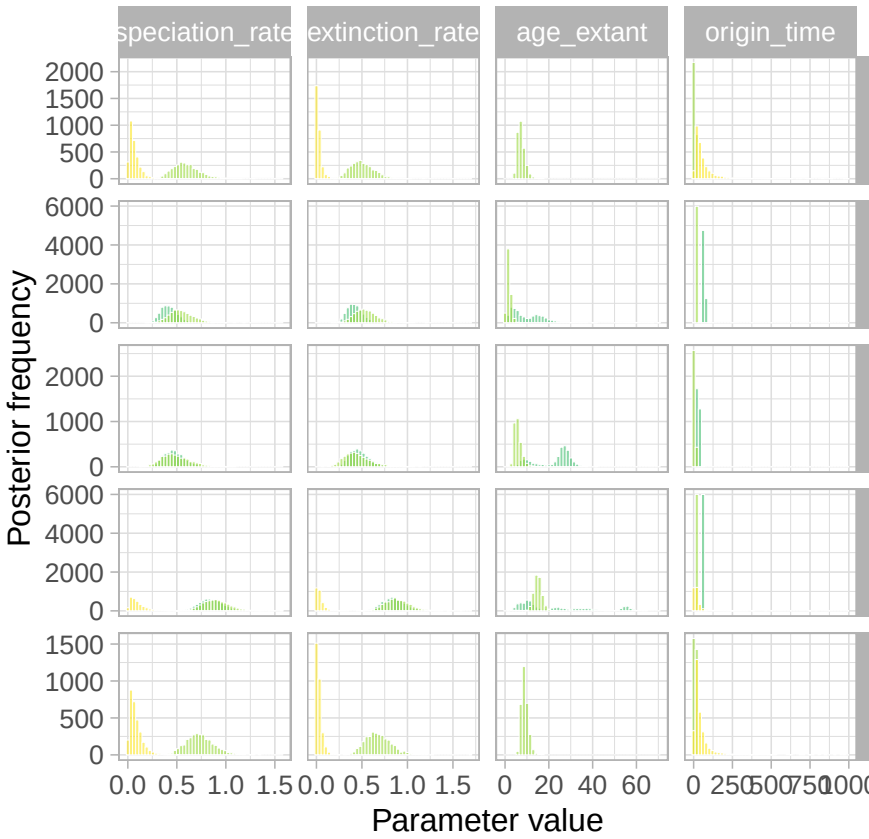
Supporting Material



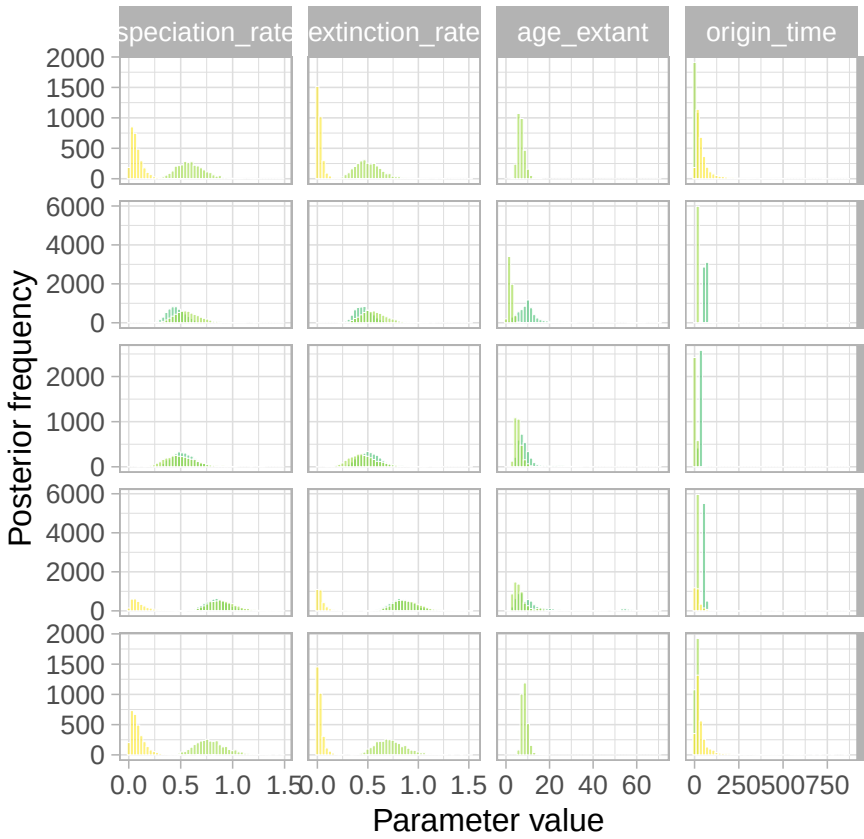
How sensitive are diversification rates to alignment strategies?



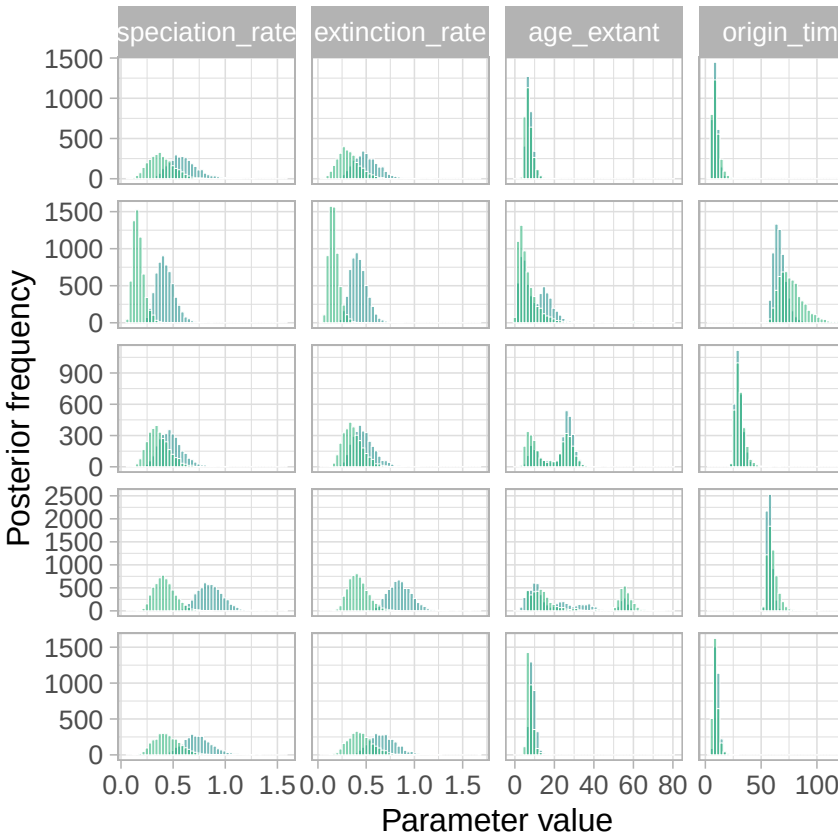
391 Sensitivity to clock model under MAFFT-agnostic



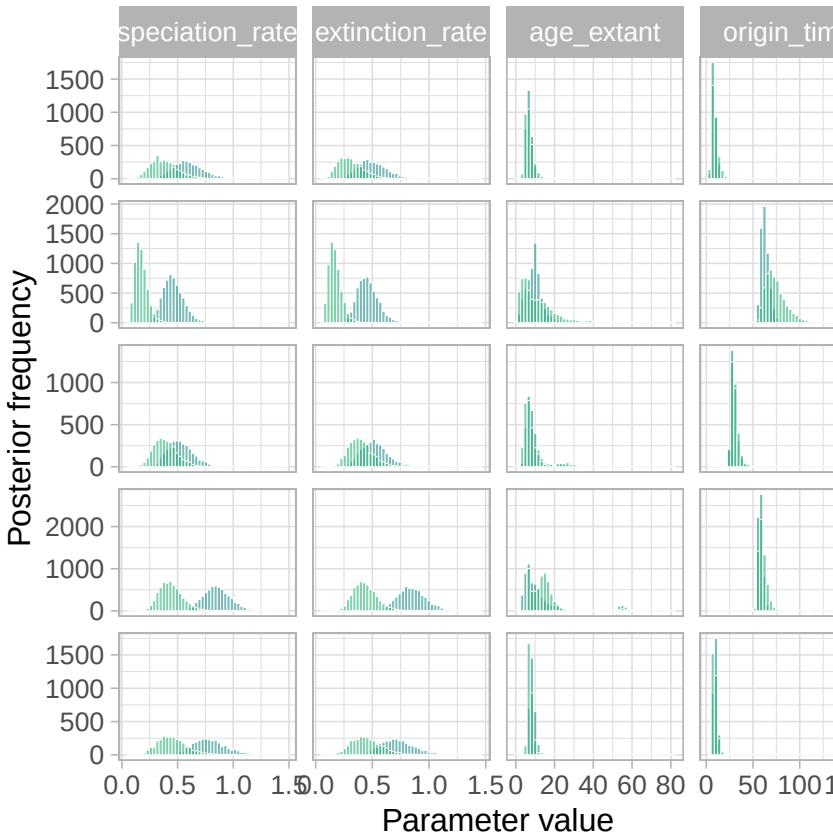
Sensitivity to fossil record under MAFFT-optimal



393 Sensitivity to fossil record under MAFFT-agnostic



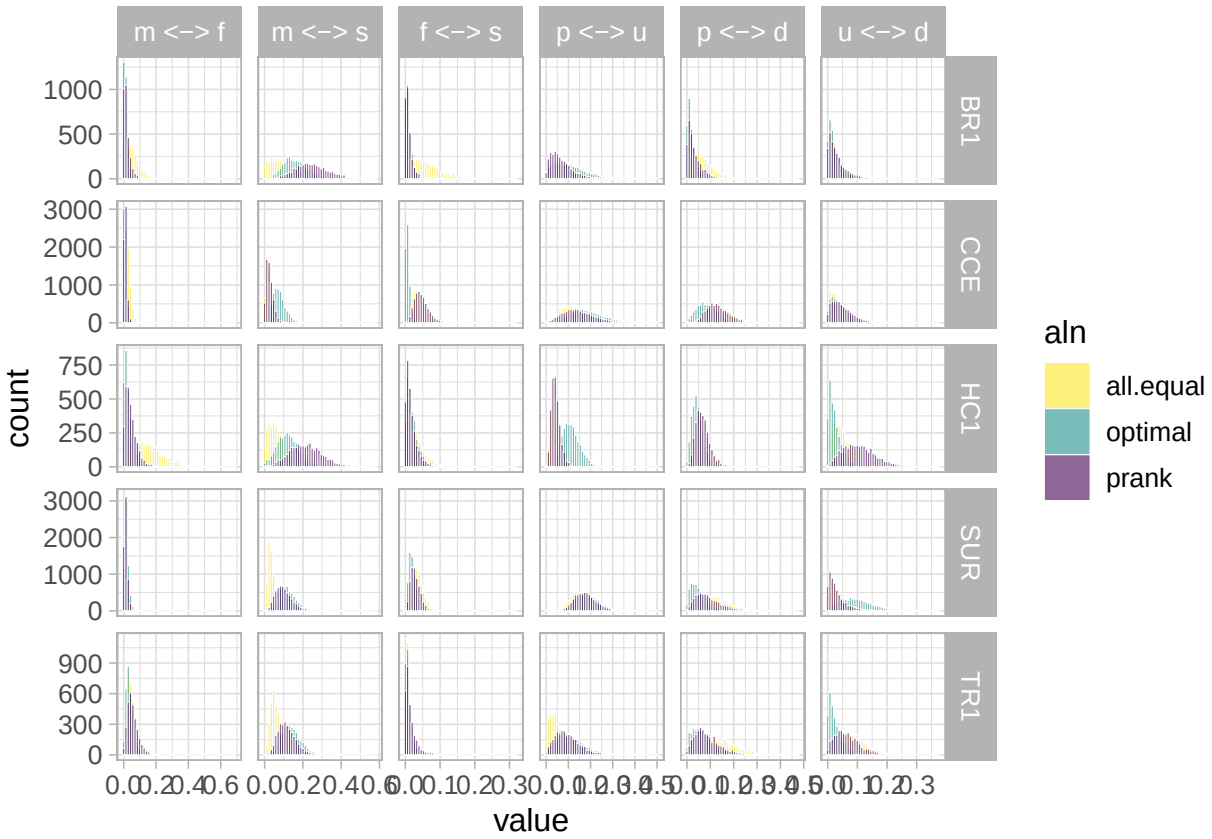
394 Sensitivity to fossil sampling under MAFFT-optimal



Sensitivity to fossil sampling under MAFFT-agnostic

Substitution rates within sound classes

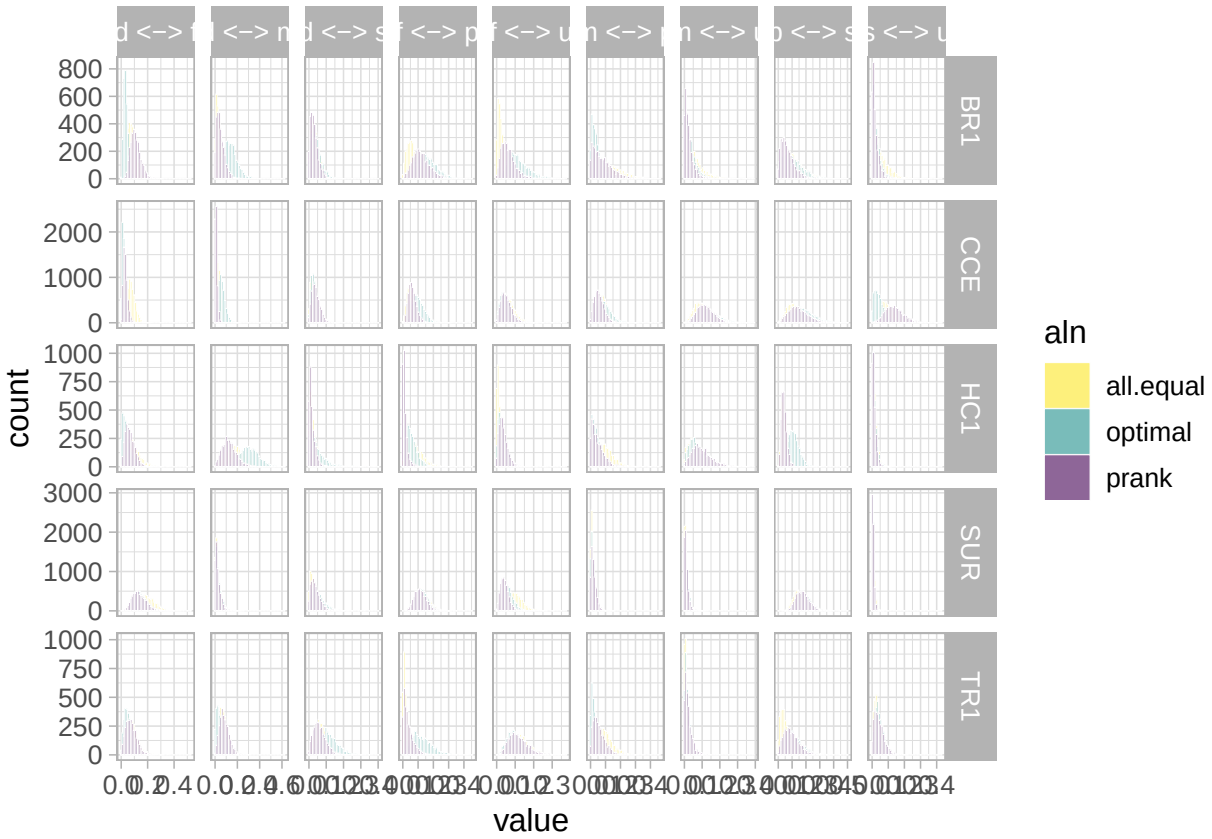
Assuming a relaxed clock, sampling all historical records and using old fossils (gapped historical record) when



available.

Substitution rates between sound classes

Assuming a relaxed clock, sampling all historical records and using old fossils (gapped historical record) when



available.

References

Aplin L.M. 2019. Culture and cultural evolution in birds: A review of the evidence. Animal Behaviour. 147:179–187.

Araya-Salas M., Smith-vidaurre G., Mennill D.J., Cahill J., Gonzalez-Gomez P.L., Wright T.F. 2019. Social group signatures in hummingbird displays provide evidence of co-occurrence of vocal and visual learning. Proceedings of the Royal Society B: Biological Sciences. 286.

Araya-Salas M., Wright T. 2013. Open-ended song learning in a hummingbird. Biology Letters. 9:20130625.

Bapteste E., Iersel L. van, Janke A., Kelchner S., Kelk S., McInerney J.O., Morrison D.A., Nakhleh L., Steel M., Stougie L., others. 2013. Networks: Expanding evolutionary thinking. Trends in Genetics. 29:439–441.

Bentley R.A., Hahn M.W., Shennan S.J. 2004. Random drift and culture change. Proceedings of the Royal Society of London. Series B: Biological Sciences. 271:1443–1450.

Boyd R., Richerson P.J. 1985. Culture and the Evolutionary Process. Chicago: The University of Chicago Press.

Bromham L., Duchêne S., Hua X., Ritchie A.M., Duchêne D.A., Ho S.Y.W. 2018. Bayesian molecular dating:

Opening up the black box. Biological Reviews. 93:1165–1191.

Bürkner P.-C. 2017. Bayesian Distributional Non-Linear Multilevel Modeling with the R Package brms. arXiv.:1705.11123.

Caetano D.S., Beaulieu J.M. 2020. Comparative analyses of phenotypic sequences using phylogenetic trees. The American Naturalist. 195:E38–E50.

Catchpole C.K., Slater P.J.B. 2003. Bird song: Biological themes and variations. Cambridge: Cambridge University Press.

Cavalli-Sforza L.L., Feldman M.W. 1981. Cultural transmission and evolution: A quantitative approach. Princeton: Princeton University Press.

Chatzou M., Magis C., Chang J.-M., Kemena C., Bussotti G., Erb I., Notredame C. 2016. Multiple sequence alignment modeling: Methods and applications. Briefings in Bioinformatics. 17:1009–1023.

Collard M., Shennan S.J., Tehrani J.J. 2006. Branching, blending, and the evolution of cultural similarities and differences among human populations. Evolution and Human Behavior. 27:169–184.

Darwin C. 1871. The descent of man and selection in relation to sex. J. Murray.

Garland E.C., Rendell L., Lamoni L., Poole M.M., Noad M.J. 2017. Song hybridization events during revolutionary song change provide insights into cultural transmission in humpback whales. Proceedings of the National Academy of Sciences. 114:7822–7829.

Gavryushkina A., Heath T.A., Ksepka D.T., Stadler T., Welch D., Drummond A.J. 2017. Bayesian total-evidence dating reveals the recent crown radiation of penguins. Systematic biology. 66:57–73.

Gavryushkina A., Welch D., Stadler T., Drummond A.J. 2014. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. PLoS Computational Biology. 10:e1003919.

Gjesfjeld E., Chang J., Silvestro D., Kelty C., Alfaro M. 2016. Competition and extinction explain the evolution of diversity in American automobiles. Palgrave Communications. 2:1–6.

Gjesfjeld E., Silvestro D., Chang J., Koch B., Foster J.G., Alfaro M.E. 2020. A quantitative workflow for modeling diversification in material culture. PloS one. 15:e0227579.

Gray R.D., Greenhill S.J., Ross R.M. 2007. The pleasures and perils of Darwinizing culture (with phylogenies). Biological Theory. 2:360–375.

Heath T.A., Huelsenbeck J.P., Stadler T. 2014. The fossilized birth–death process for coherent calibration of divergence-time estimates. Proceedings of the National Academy of Sciences. 111:E2957–E2966.

Höhna S., Coghill L.M., Mount G.G., Thomson R.C., Brown J.M. 2018. P3: Phylogenetic posterior prediction in RevBayes. Molecular Biology and Evolution. 35:1028–1034.

Höhna S., Landis M.J., Heath T.A., Boussau B., Lartillot N., Moore B.R., Huelsenbeck J.P., Ronquist F. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification

language. Systematic Biology. 65:726–736.

Holland S.M. 2016. The non-uniformity of fossil preservation. Philosophical Transactions of the Royal Society B: Biological Sciences. 371:20150130.

Jesmer B.R., Merkle J.A., Goheen J.R., Aikens E.O., Beck J.L., Courtemanch A.B., Hurley M.A., McWhirter D.E., Miyasaki H.M., Monteith K.L., others. 2018. Is ungulate migration culturally transmitted? Evidence of social learning from translocated animals. Science. 361:1023–1025.

Katoh K., Misawa K., Kuma K., Miyata T. 2002. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research. 30:3059–3066.

Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution. 30:772–780.

Kempe M., Lycett S., Mesoudi A. 2012. An experimental test of the accumulated copying error model of cultural mutation for Acheulean handaxe size. PLoS One. 7:e48333.

Kershenbaum A., Blumstein D.T., Roch M.A., Akçay C., Backus G., Bee M.A., Bohn K., Cao Y., Carter G., Cäsar C., Coen M., Deruiter S.L., Doyle L., Edelman S., Ferrer-i-Cancho R., Freeberg T.M., Garland E.C., Gustison M., Harley H.E., Huetz C., Hughes M., Hyland Bruno J., Ilany A., Jin D.Z., Johnson M., Ju C., Karnowski J., Lohr B., Manser M.B., Mccowan B., Mercado E., Narins P.M., Piel A., Rice M., Salmi R., Sasahara K., Sayigh L., Shiu Y., Taylor C., Vallejo E.E., Waller S., Zamora-Gutierrez V., Akçay Ç., Backus G., Bee M.A., Bohn K., Cao Y., Carter G., Cäsar C., Coen M., Deruiter S.L., Doyle L., Edelman S., Ferrer-i-Cancho R., Freeberg T.M., Garland E.C., Gustison M., Harley H.E., Huetz C., Hughes M., Hyland Bruno J., Ilany A., Jin D.Z., Johnson M., Ju C., Karnowski J., Lohr B., Manser M.B., Mccowan B., Mercado E., Narins P.M., Piel A., Rice M., Salmi R., Sasahara K., Sayigh L., Shiu Y., Taylor C., Vallejo E.E., Waller S., Zamora-Gutierrez V., Akçay C., Backus G., Bee M.A., Bohn K., Cao Y., Carter G., Cäsar C., Coen M., Deruiter S.L., Doyle L., Edelman S., Ferrer-i-Cancho R., Freeberg T.M., Garland E.C., Gustison M., Harley H.E., Huetz C., Hughes M., Hyland Bruno J., Ilany A., Jin D.Z., Johnson M., Ju C., Karnowski J., Lohr B., Manser M.B., Mccowan B., Mercado E., Narins P.M., Piel A., Rice M., Salmi R., Sasahara K., Sayigh L., Shiu Y., Taylor C., Vallejo E.E., Waller S., Zamora-Gutierrez V., Akçay Ç., Backus G., Bee M.A., Bohn K., Cao Y., Carter G., Cäsar C., Coen M., Deruiter S.L., Doyle L., Edelman S., Ferrer-i-Cancho R., Freeberg T.M., Garland E.C., Gustison M., Harley H.E., Huetz C., Hughes M., Hyland Bruno J., Ilany A., Jin D.Z., Johnson M., Ju C., Karnowski J., Lohr B., Manser M.B., Mccowan B., Mercado E., Narins P.M., Piel A., Rice M., Salmi R., Sasahara K., Sayigh L., Shiu Y., Taylor C., Vallejo E.E., Waller S., Zamora-Gutierrez V. 2016. Acoustic sequences in non-human animals: A tutorial review and prospectus. Biological Reviews. 91:13–52.

Kidwell S.M., Holland S.M. 2002. The quality of the fossil record: Implications for evolutionary analyses.

Annual Review of Ecology and Systematics. 33:561–588.

Laland K.N., Hoppitt W. 2003. Do animals have culture? Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews. 12:150–159.

Laland K.N., Williams K. 1997. Shoaling generates social learning of foraging information in guppies. Animal Behaviour. 53:1161–1169.

Ligon R.A., Diaz C.D., Morano J.L., Troscianko J., Stevens M., Moskeland A., Laman T.G., Scholes E. 2018. Evolution of correlated complexity in the radically different courtship signals of birds-of-paradise. PLOS Biology. 16:e2006962.

Löytynoja A. 2012. Alignment Methods: Strategies, Challenges, Benchmarking, and Comparative Overview. In: Anisimova M., editor. Evolutionary genomics: Statistical and computational methods, volume 1. Totowa, NJ: Humana Press. p. 203–235.

Löytynoja A., Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proceedings of the National Academy of Sciences of the United States of America. 102:10557–10562.

Löytynoja A., Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science. 320:1632–1635.

Luncz L.V., Boesch C. 2014. Tradition over trend: Neighboring chimpanzee communities maintain differences in cultural behavior despite frequent immigration of adult females. American Journal of Primatology. 76:649–657.

Lunter G., Miklós I., Drummond A., Jensen J.L., Hein J. 2005. Bayesian coestimation of phylogeny and sequence alignment. Bmc Bioinformatics. 6:1–10.

Luo A., Duchêne D.A., Zhang C., Zhu C.-D., Ho S.Y.W. 2020. A simulation-based evaluation of tip-dating under the fossilized birth–death process. Systematic Biology. 69:325–344.

Lutzoni F., Wagner P., Reeb V., Zoller S. 2000. Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology. Systematic Biology. 49:628–651.

Mesoudi A. 2017. Pursuing Darwin's curious parallel: Prospects for a science of cultural evolution. Proceedings of the National Academy of Sciences. 114:7853–7860.

Morlon H. 2014. Phylogenetic approaches for studying diversification. Ecology Letters. 17:508–525.

Payne R.S., McVay S. 1971. Songs of humpback whales. Science. 173:585–597.

Perreault C. 2012. The pace of cultural evolution. PLoS One. 7:e45150.

Philippe H., Douady C.J. 2003. Horizontal gene transfer and phylogenetics. Current Opinion in Microbiology. 6:498–505.

Plummer M., Best N., Cowles K., Vines K. 2006. CODA: Convergence Diagnosis and Output Analysis for MCMC. R News. 6:7–11.

R Core Team. 2021. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rama T. 2018. Three tree priors and five datasets: A study of indo-european phylogenetics. Language Dynamics and Change. 8:182–218.

Redelings B.D., Suchard M.A. 2005. Joint Bayesian estimation of alignment and phylogeny. Systematic biology. 54:401–418.

Ritchie A.M., Ho S.Y.W. 2019. Influence of the tree prior and sampling scale on bayesian phylogenetic estimates of the origin times of language families. Journal of Language Evolution. 4:108–123.

Rivera-Cáceres K.D., Quirós-Guerrero E., Araya-Salas M., Searcy W.A. 2016. Neotropical wrens learn new duet rules as adults. Proceedings of the Royal Society B: Biological Sciences. 283.

Sagart L., Jacques G., Lai Y., Ryder R.J., Thouzeau V., Greenhill S.J., List J.-M. 2019. Dated language phylogenies shed light on the ancestry of Sino-Tibetan. Proceedings of the National Academy of Sciences. 116:10317–10322.

Stadler T., Yang Z. 2013. Dating phylogenies with sequentially sampled tips. Systematic biology. 62:674–688.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123:585–595.

Warnow T. 2021. Revisiting Evaluation of Multiple Sequence Alignment Methods. Multiple sequence alignment. Springer. p. 299–317.

Watterson G.A. 1975. On the number of segregating sites in genetical models without recombination. Theoretical population biology. 7:256–276.

Whiten A., Horner V., De Waal F.B.M. 2005. Conformity to cultural norms of tool use in chimpanzees. Nature. 437:737–740.

Williams H., Levin I.I., Norris D.R., Newman A.E.M., Wheelwright N.T. 2013. Three decades of cultural evolution in Savannah sparrow songs. Animal Behaviour. 85:213–223.

Xie W., Lewis P.O., Fan Y., Kuo L., Chen M.-H. 2010. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Systematic Biology. 60:150–160.

Yang Z., Rannala B. 2012. Molecular phylogenetics: Principles and practice. Nature Reviews Genetics. 13:303–314.

Zhang C., Stadler T., Klopfstein S., Heath T.A., Ronquist F. 2016. Total-evidence dating under the fossilized birth–death process. Systematic Biology. 65:228–249.

Zhang H., Ji T., Pagel M., Mace R. 2020. Dated phylogeny suggests early neolithic origin of sino-tibetan languages. Scientific reports. 10:1–8.