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@zoologi.su.se)[3,4*]

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

2026-02-27

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

[2] Lab of Ornithology, Cornell University

[3] Department of Zoology, Stockholm University, Stockholm 106-91, Sweden

[4] Department of Biological Sciences, National University of Singapore, Singapore 117558, Singapore

[5] Department of Biology, University of Washington

* Both authors contributed equally

† To whom correspondence should be addressed

Keywords: acoustic sequences, Long-billed Hermit, model adequacy, node dating, phylogenetics, posterior predictive simulation, sequence alignment

Abstract

Introduction

The idea that culture changes and diversifies 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 other 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). In the last century, the formal modelling of cultural change as evolution, whether in humans or non-human animals, has been conducted by drawing analogies between cultural and population genetic processes (Cavalli-Sforza and Feldman 1981; Boyd and Richerson 1985), or between cultural and phylgenetic diversification (Gray et al. 2007; Mesoudi 2017). 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. tool design vs. language). In order to address the long-standing question of whether cultural change 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.

Some learnt 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 facilitates 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. Estimating the absolute fit of such models to cultural data is crucial to evaluate the utility of phylogenetic approaches for our understanding of cultural evolution and diversification.

Cultural evolution poses special 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. Such an approach can help us determine whether the analogy between organic evolution and cultural change holds for other cultural phenomena. Because bird songs are often socially learnt and linearly composed of discrete subunits, they can be used to examine the suitability of the FBDP as a phylogenetic model of cultural diversification.

The Long-billed Hermit (*Phaethornis longirostris*; Fig. 1a) produces songs that can be represented as sequences of discrete sounds fused together into an unbroken signal (Fig. 1b; 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). Evidence of social learning in this species (*sensu* Ten Cate 2021), 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). 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 and Wolf 1979), 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. Such independence across leks provides an unmatched opportunity to investigate the robustness of phylogenetic models across different iterations of an underlying cultural diversification process.

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,parameter estimation and the fit of alternative clock models to long-billed hermit song data.

Figure 1. Socially transmitted songs in the Long-billed Hermit. **a)** A male Long-billed Hermit. **b)** Spectrograms of two songs from different males in the SUR lek, sampled in 2019. The *colour* arrow shows a pure tone and the *another colour* arrow show a vibratory sound. **c)** Locations of the study leks in the Caribbean lowlands of Costa Rica. **d)** Historical sampling of song records in each lek.

Methods

Data collection and song structure coding

Sound recordings of Long-billed Hermits were registered from 2008 to 2019, in five leks, distributed across four sites in the Caribbean slope of Costa Rica: La Selva Biological Station (leks SUR and CCE), Finca las Brisas (BR1), Hitoy Cerere Biological Reserve (HC1) and La Tirimbina Lodge (TR1) (Fig. 1c). We also included historical recordings available for three of the studied leks at La Selva and Hitoy Cerere (Stiles and Wolf 1979). 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 structure, see below) is not affected by the recording equipment in a detectable manner.

Long-billed Hermit songs are composed of two basic sound types: tonal and vibratory sounds (trills). 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) (Fig. 1b). 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. Songs were split into 20 equal-length segments, and each segment was assigned to one of these six categories, based upon visual inspection of spectrograms (Fig. S1). The choice of 20 segments per song captured a compromise between our ability do discriminate sound types, which increased with segment length, and the probability that a single sound type occurred in each segment, which decreased with segment length (Supporting Text 1). We validate this choice by assessing inter-observer repeatability of sound classification (Supporting Text 2; Fig. S2).

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 data set. 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 characters and 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 hummingbirds 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 sounds or *vice versa*. 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 made two further assumptions, namely that cultural evolution is independent between leks and

also between individuals within leks (see Discussion). If insertions and deletions of song segments occur independently among individuals, alignment length should increase as sequences are more distantly related (Löytynoja 2012). We would thus expect longer alignments in data sets composed of 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 for the same number of sequences originating from the same lek. 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 support customized alphabets and substitution-rate matrices. To use PRANK, we assumed that pure tones can be treated as ambiguous between upward and downward tones, and medium-speed trills can similarly be treated and ambiguous between fast and slow trills. We therefore used IUPAC ambiguity codes 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, the 'PRANK-TN93' alignment explicitly assumed a higher transition rate within vibratory and tonal 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 and 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 represent lineages that 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. This 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 data set of historical records of songs in hummingbird leks has three advantages in comparison to most fossil data sets 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 hummingbirds are actively displaying their calls (Stiles and Wolf 1979), we can assume detection is nearly perfect and thus there is no missing taxa in any of the sampled years.

A possible complication in our analysis is that long gaps without sampling are interspaced in the three leks with deeper historical records (HC1, CCE and SUR). 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 data set, including long gaps without lineage sampling, and with the more recent and continuously sampled data. Finally, to investigate the general impact of sampling historical records on phylogenetic inference of song diversification, we the conducted an additional set of analyses, including only terminal tips (i.e. songs observed in the last year of sampling). 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 from 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.

Phylogenetic analyses were conducted with all three alignment strategies (MAFFT-agnostic, MAFFT-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.

We conducted two independent MCMC runs for each analyses, with 150 000 iterations and an additional 15 000 of burn-in and parameter tuning every 200. 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 iteration. We assessed MCMC performance using the package *coda* v. 0.19-4 (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, as well as the multivariate estimate. We also inspected autocorrelations between draws (targeted below 0.1) and effective sample sizes (targeted above 200) for all model variables. We plotted trees for visual inspection using the package *ggtree* (Yu et al. 2017) and our general results using the package *ggplot2* (Wickham 2016) in R.

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 data sets based on the inference model. We specified a thinning of 2 iterations for the stochastic variable trace, thus simulating 3 000 datasets for the 'large' leks (CCE, SUR) and 1 500 datasets for the 'small' leks (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, used to calculate Tajima's D, and a measure of character diversity. 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 lower 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 exactly 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, the posterior predictive p-value was set to 1.

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 and Foulds 1981) with the R package *phangorn* v.2.11.1 (Schliep 2011). Topological distances were projected in a bidimensional space using Classic Multidimentional Scalling in order to quantify topological space.

Only tree tips pairwise shared between trees were considered when calculating distances. We estimated the overall spread of the topological space (i.e. space size) for different models as a metric of within-model topological congruence. Topological space size was quantified as the 95% kernel density area. 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 intersection-over-union of the two 95% kernel density spaces (i.e. proportion of the joint area of two spaces that was shared). Topological congruence descriptors were calculated using the R package PhenotypeSpace v.0.1.0 (Araya-Salas and Odom 2022).

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 (modeled with a gaussian distribution), model alignment strategy, use of historical data, historical record completeness, and clock model as predictors and lek as a varying intercept effect (i.e. random effect). Regression models were run in Stan (Carpenter et al. 2017) through the R platform (R Core Team 2021) using the package *brms* v.2.22.0 (Bürkner 2017). We also computed multiple comparisons of alignment strategies (similar to post hoc tests in frequentist statistics) using the joint posterior distribution of the model parameters. We present effect size point estimates as median posterior estimates, and report 95% highest posterior density intervals (HPDIs) as a measure of uncertainty. We evaluated the predictive performance of all models, measured by the LOO Information Criterion (LOOIC; Vehtari et al. 2017), by comparing each model against its correspondent null model (i.e., a model with no fixed effects). Models were compared using the expected log predictive density (ELPD). Models were run on three chains for 2500 iterations, following a warm-up of 2500 iterations using weakly informative priors. Effective sample size was kept above 3000 for all parameters. Potential scale reduction factor was used to assess model convergence and kept below 1.05 for all parameter estimates. 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. Posterior predictive check plots were also used to inspect if model predictions aligned with the distribution and variability of the observed data.

Finally, we 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, extinction, and net diversification rates, as well as the age of the MRCA of all songs (hereafter 'root age', including extinct lineages) and the age of the MRCA of only extant songs (hereafter 'crown age', including song present in the last year of sampling). Following Muff et al. (2021), we communicate our statistical results in the language of evidence.

Results

We used the FBDP to model cultural change and diversification in five independent leks of the Long-billed Hermit. These natural leks differ markedly in size and in the length and completeness of their historical sampling record (Fig. 1d; 2). MCMC model diagnostics reflect this disparity. The MCMC analysis of the smallest lek (BR1) under the global clock model was the only scenario in which all model parameters met out mixing (autocorrelation < 0.1) and convergence (psrf < 1.05) criteria (Table S1). In larger leks and more complex models (i.e. relaxed clock) more model parameters, especially branch rate parameters, failed the mixing, and to a lesser extent, the convergence criteria (Table S1). Nonetheless, in all but one analysis (SUR lek using the MAFFT-optimal alignment and all observations of all historical records and a global clock), estimation of the most relevant model parameters (likelihood, root age, crown age and diversification rates) converged between independent runs (Table S1). We therefore used these models to investigate absolute model fit, relative fit of alternative clock models, and the effects of alignment strategies and the historical record on topological convergence and diversification inferences. We return to the caveats in parameter estimation in the Discussion.

Figure 2. Maximum *a posteriori* (MAP) trees for each lek under a relaxed clock model and utilizing the entire historical song record and a phylogenetically informed alignment algorithm (PRANK). **a)** BR1, **b)** TR1, **c)** HC1, **d)** CCE, **e)** SUR.

Clock model selection

A relaxed clock model of song evolution (in which different song lineages evolve at different rates) was generally supported, but the strength of this support depended on the historical record and sampling strategy. When historical data was entirely excluded, there was no increase in ML by relaxing the clock model (Fig 3). 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, were incorporated in the macroevolutionary process (Fig 3). While this trend was present in most leks and data sets, it tended to be stronger under the phylogenetically-informed PRANK-TN93 alignment, especially in the historically largest lek (SUR). Hereafter, we present the results of analyses using the relaxed clock model.

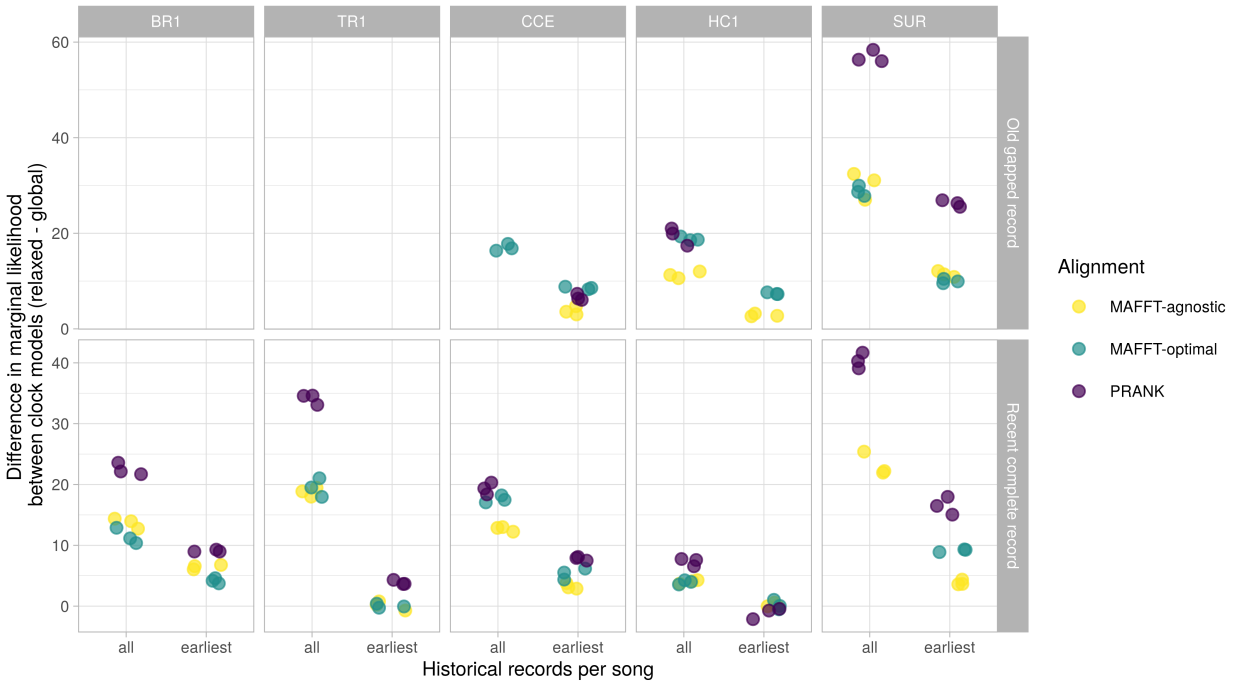

Figure 3. Relative fit of relaxed and global clock models in phylogenetic inference of song evolution under the FBDP. The fit of alternative clock models was compared in five leks (BR1, TR1, HC1, CCE, and SUR) using marginal likelihood approximation. Two leks (BR1 and TR1) lacked deep historical data (no sampling "including oldest records") and three leks contained enough extant songs to estimate clock model fit using a BDP (without historical data).

Model reliabilty

Simulated data under our inference models accurately reflected certain features of most empirical data sets. Namely, the number of segregating and invariant sites, the maximum length of variable and invariant blocks and the minimum pairwise distance between sequences were generally in agreement between simulated and empirical data (Table 1; Fig. 4; Fig. S3-S32). The two measures of song element diversity (Watterson's θ

and π), analogous to population genetic diversity, were reliably modeled with the exception of about a third of the PRANK alignments, and four MAFFT-agnostic alignments, in which simulated data underestimated diversity (Fig. 5; Fig. S23-S29). Consequently, several of these alignments also resulted in relatively low values of Tajima's D in simulated data, consistent with our models overestimating the effect of drift (Fig. S30-32). Finally, the the maximum pairwise distance between sequences was the data feature most often in conflict between simulated and empirical data sets (Fig. 6), and overall underestimated in simulated data (Fig. S20-S22).

Discrepancies between empirical and simulated data were found in all leks and alignment strategies (Fig. S3-S32). For the same lek and historical record, PRANK-TN93 alignments tended to result in relatively high song-element diversity (π), in both empirical and simulated data sets. Nonetheless, such high song diversity was often underestimated, while the strength of drift in these alignments was overestimated, particularly in the TR1 lek (Fig. 5; Fig. S27-32). Similarly, where estimates of the maximum pairwise distance between songs were highest, mainly in PRANK-TN93 and MAFFT-optimal alignments, such song divergence tended to be underestimated (Fig. 6; Fig. S20-22).

Table 1. Summary of absolute model fit statistics obtained by comparing properties of empirical data to posterior predictive data simulations.

Alignment	Historical record	Mean PPE	SD	Min	Max	Percent P-value > 0.05
MAFFT-agnostic	Old gapped	0.63	0.79	0	3.23	0.97
MAFFT-agnostic	Recent complete	0.82	0.86	0	3.94	0.91
MAFFT-agnostic	None	0.53	0.55	0	2.10	1.00
MAFFT-optimal	Old gapped	0.95	0.84	0	3.09	0.95
MAFFT-optimal	Recent complete	0.85	0.83	0	3.93	0.88
MAFFT-optimal	None	1.05	0.74	0	2.89	0.96
PRANK-TN93	Old gapped	1.00	0.87	0	4.37	0.92
PRANK-TN93	Recent complete	0.81	0.88	0	4.06	0.93
PRANK-TN93	None	1.36	1.04	0	3.68	0.70

Treespace congruence

The spread of the topological space increased when including all historical records (effect size: 0.031, uncertainty interval (UI): 0.023 - 0.040) and when including fossils (effect size: 0.012, UI: 0.005 - 0.019), but

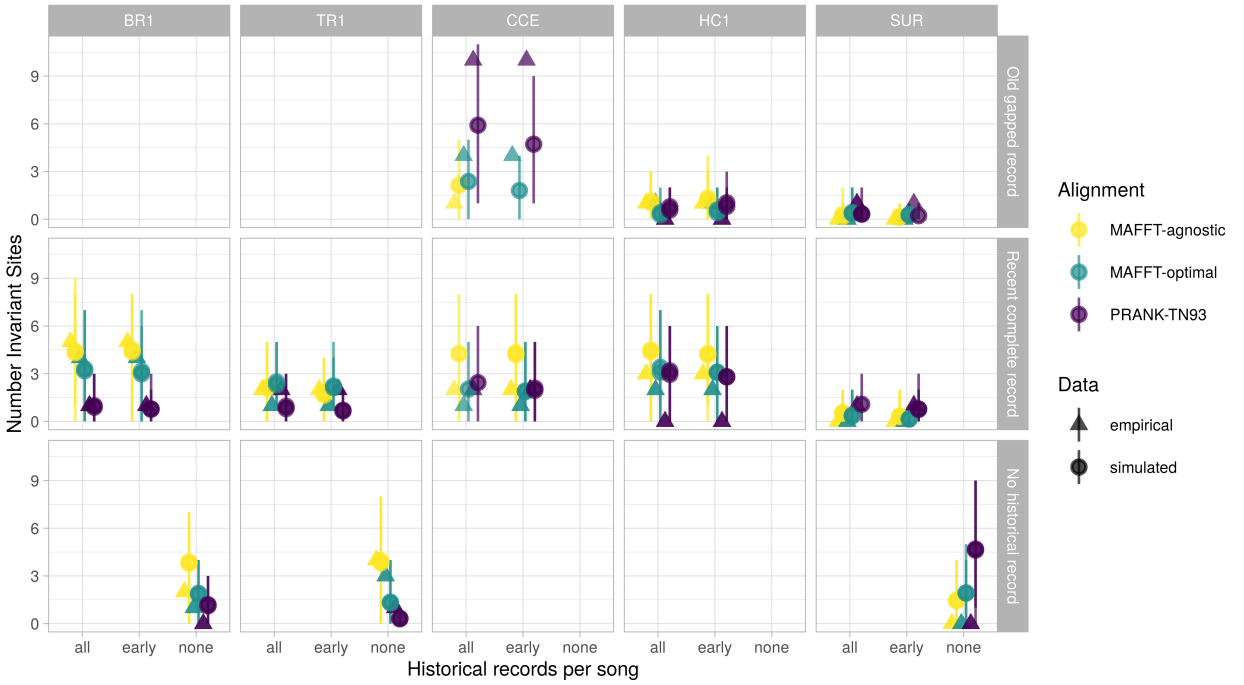

Figure 4. Number of invariant sites in empirical song data and posterior predictive data simulations (see Methods). For simulated data, we report the mean and 95% HPD interval of 1500 (BR1, TR1, HC1) or 3000 (CCE and SUR) simulations.

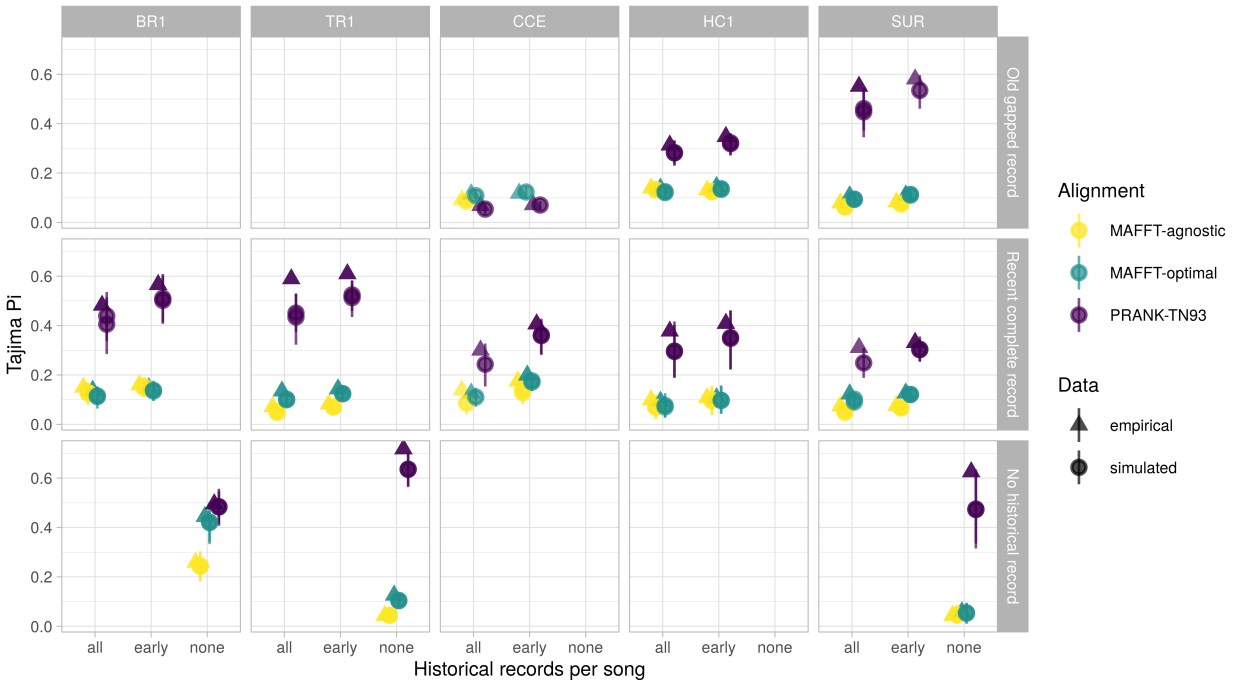

Figure 5. Song element diversity (π) in empirical song data and posterior predictive data simulations (see Methods). For simulated data, we report the mean and 95% HPD interval of 1500 (BR1, TR1, HC1) or 3000 (CCE and SUR) simulations.

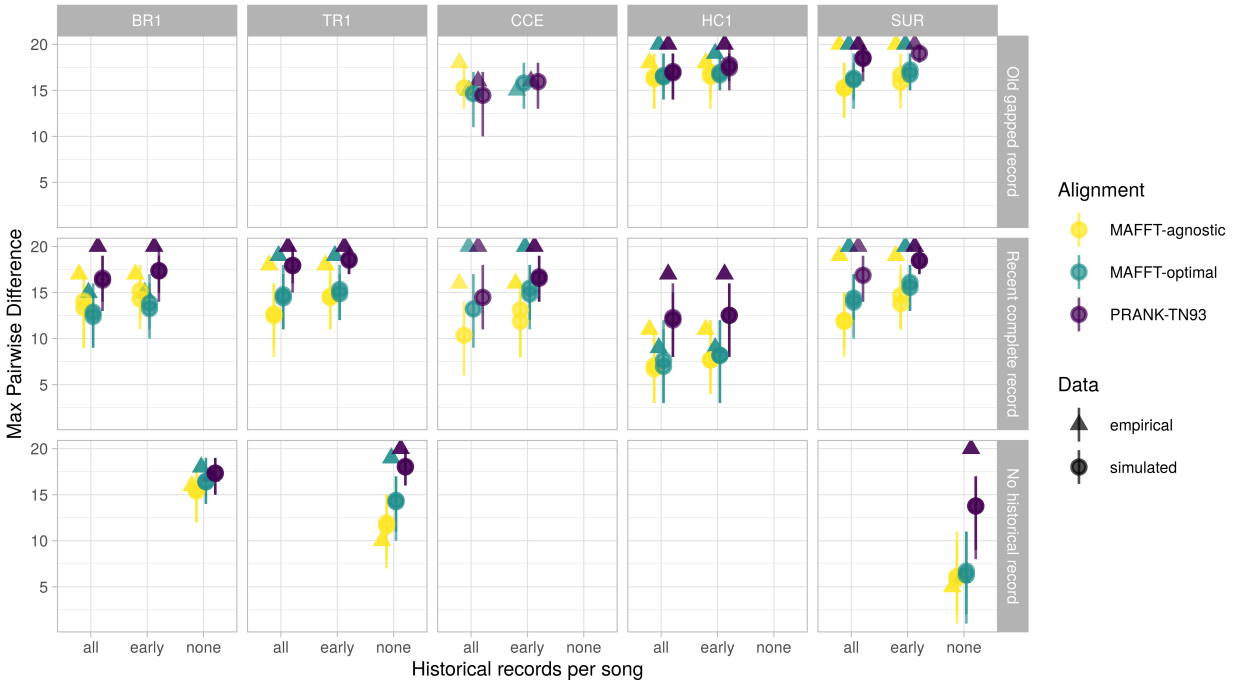

Figure 6. Maximum pairwise distance between songs in empirical song data and posterior predictive data simulations (see Methods). For simulated data, we report the mean and 95% HPD interval of 1500 (BR1, TR1, HC1) or 3000 (CCE and SUR) simulations.

was not affected by the use of different clock models (effect size: -0.002, UI: -0.009 - 0.005). For the alignment strategy only the MAFFT-optimal alignment increased the spread of the topological space compared to the PRANK-TN93 alignment (effect size: -0.009, UI: -0.018 - 0.000) no other differences in space size were detected between alignment strategies (PRANK-TN93 VS MAFFT-agnostic, effect size: -0.005, UI:-0.014 - 0.003; MAFFT-optimal VS MAFFT-agnostic, effect size: 0.004, UI: -0.005 - 0.013). The dissimilarity of the topological space between models increased by the inclusion of all historical records (effect size: 0.174, UI: 0.130 - 0.224) and fossils (effect size: 0.176, UI: 0.119 - 0.235), but not by the use of different clock models (effect size: 0.030, UI: -0.015 - 0.076). The PRANK-TN93 alignment increased topological space dissimilarity when compared to both MAFFT-optimal (effect size: 0.102, UI: 0.045 - 0.159) and MAFFT-agnostic alignments (effect size: 0.110, UI: 0.043 - 0.180), but did not differ between the MAFFT alignments (effect size: 0.008, UI: -0.049 - 0.067).

Diversification dynamics

Diversification rates and divergence times were 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 speciation and extinction rates and markedly uncertain root ages, in comparison to analyses including

historical songs (Fig. 7; S33-S35). For the three leks in which a long but gapped historical record could be contrasted with a shorter but complete record (CCE, HC1, and SUR), the inclusion of more ancient records resulted in older crown and root age estimates (Fig. 7; S33-S35). In contrast, strong effects of alignment strategies were found in only two leks. In the largest lek (SUR), the PRANK-TN93 alignment led to a moderate decrease in speciation and extinction rate and a relatively strong increase in divergence times (Fig. 8). In the HC1 lek both the PRANK-TN93 and MAFFT-optimal alignments caused older crown age estimates (Fig. 8), compared to the MAFFT-agnostic alignment. Finally, sampling all historical occurrences, as opposed to only the earliest occurrence of each unique song, resulted in increased speciation and extinction rates and slightly more precise estimates of divergence times (Fig. 9; Fig. S36-S38).

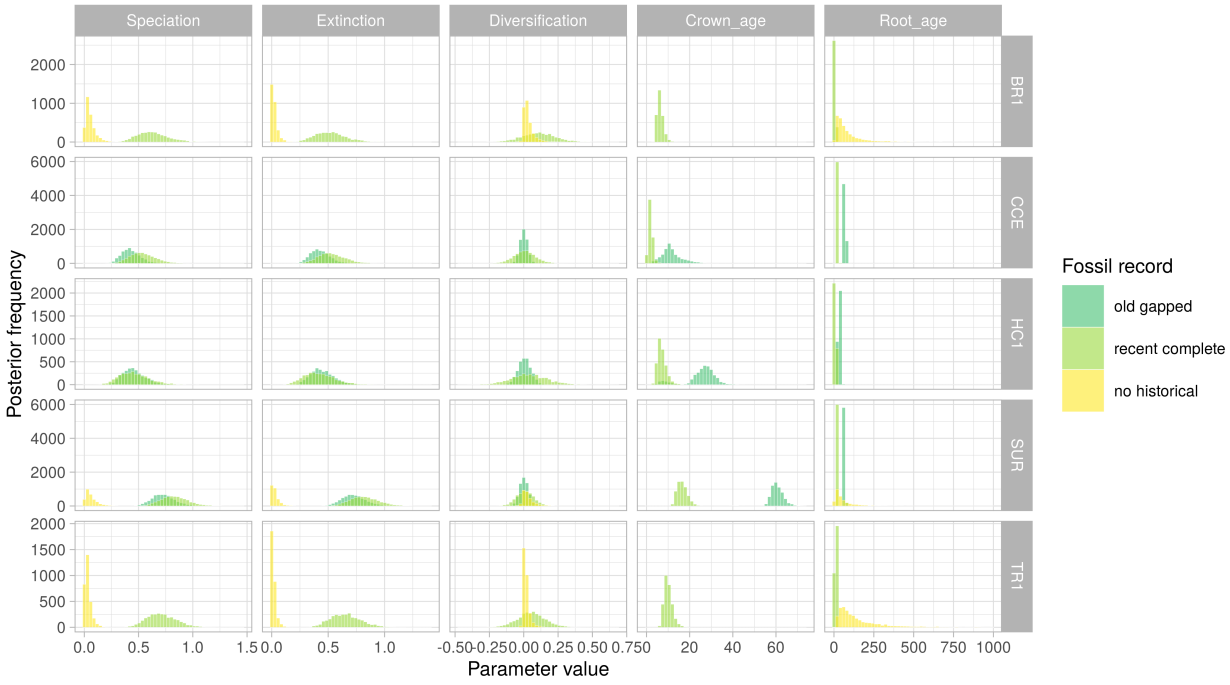

Figure 7. Age and diversification estimates under different sets of historical song records in five leks of the Long-billed Hermit. Histograms show posterior distributions of parameter estimates in models using a relaxed molecular clock and the PRANK-TN93 alignment strategy. For data sets including historical song records, all occurrences of such record are included in the analysis.

Discussion

Many researchers view cultural change as a process akin to organic evolution [REF]. However, a basic prediction of this parallel has largely evaded scrutiny. We typically do not know if evolutionary models and especially phylogenetic models applied to cultural change capture the most prominent features of this process, resulting absolute fit to cultural data [REF exceptions?]. Here, we address this question using the socially learnt songs of the Long-billed Hermit in multiple independent leks, and a phylogenetic model, the fossilized

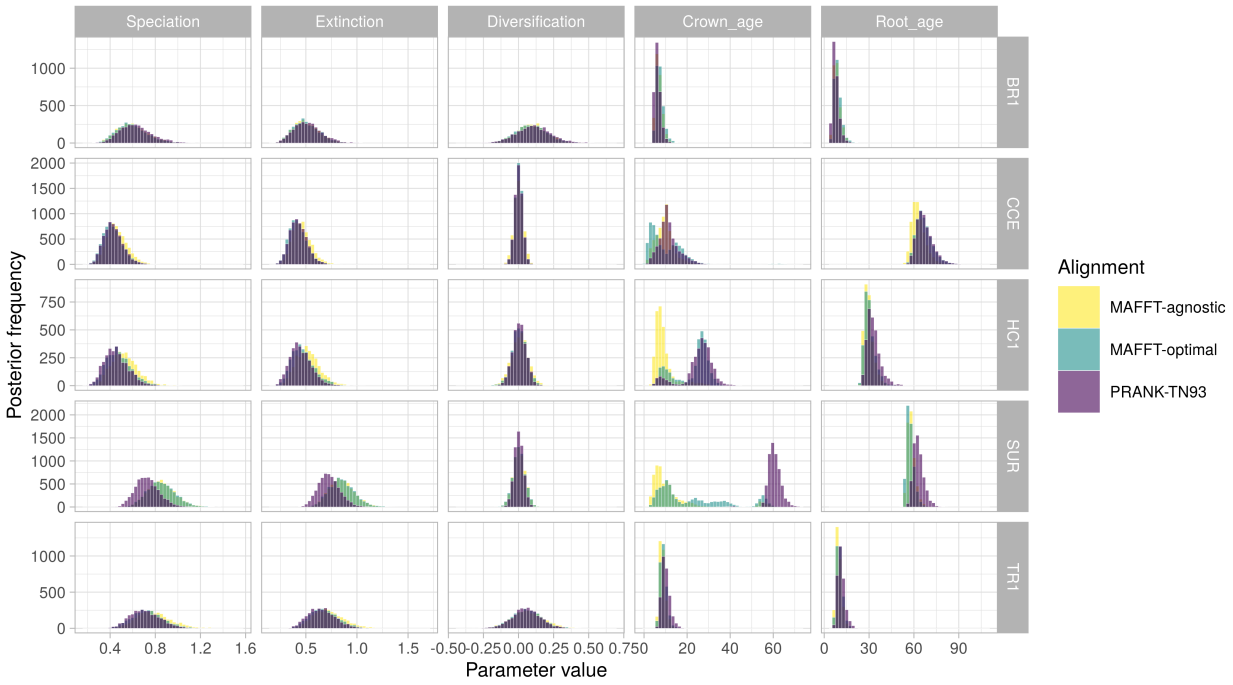

Figure 8. Age and diversification estimates under different alignment strategies in five leks of the Long-billed Hermit. Histograms show posterior distributions of parameter estimates in models using a relaxed molecular clock and all historical records. All available historical records are included in the analysis.

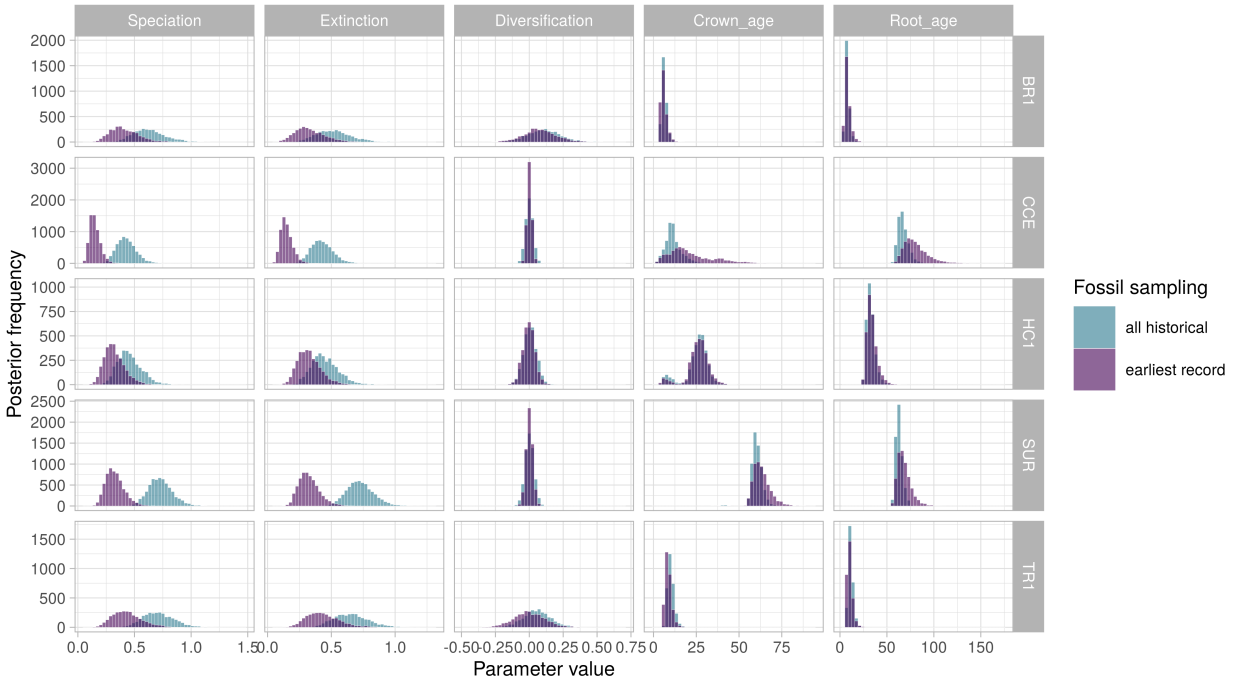

Figure 9. Age and diversification estimates of songs in five leks of the Long-billed Hermit. Histograms show posterior distributions of parameter estimates in models using a relaxed molecular clock and the PRANK-TN93 alignment strategy. We compare analyses using all occurrences of historical songs, including identical songs sampled serially, and analyses including each unique song at its earliest occurrence.

birth-death process, that is particularly suitable for serially-sampled data. We found that these phylogenetic models are overall reliable, although they tend to underestimate song element diversity and song divergence in some scenarios. We also asked how the historical song record, analogous to the fossil record, and alignment strategies, ported from molecular phylogenetics, influence parameter estimation. In line with theoretical expectations [REF Heath?], inclusion of historical data resulted in more precise root-age estimates (Fig. 7), a higher rate of lineage divergence ("speciation") and extinction (Fig. 7). Our analyses with historical song records also lent support for heterogeneous branch rates (Fig. 3). If available, older historical records prompted deeper divergence times (Fig. 7). On the other hand, alignment strategies had very strong effects on tree topology (Fig. ?), and weak to moderate effects on diversification dynamics (Fig. 8).

Posterior predictive simulations suggested that the FBDP is a plausible model for cultural change in the Long-billed Hermit. Nearly 90% of posterior predictive effect sizes (PPEs) between empirical and simulated data under the FBDP were below 2.00 (mean PPE = 0.84, sd = 0.85, min = 0, max = 4.37; Table 1). Moreover, only one scenario, in which all historical records were excluded (SUR, PRANK-TN93 alignment), produced consistent evidence for model inadequacy in data-based posterior predictive checks (Fig. S3-S32). Even though tests of model adequacy are important safeguards against estimation biases (Carstens et al. 2022), tools for their implementation in Bayesian phylogenetics are relatively new [REFs incl Hohna 2018 and some Duchene]. To our knowledge, these tools have been advocated [BEAST REf] but not yet utilized in cultural phylogenetics [but see REFs for other models of cultural evolution]. When applied to molecular and morphological phylogenetics, posterior predictive checks often reveal a wide range of model fits, depending on the locus (or trait) and clade analyzed (Richards et al. 2018; May et al. 2021; Khouri et al. 2022). We were therefore surprised to find that FBD models were similarly adequate in most of our data sets, ranging from 8 'taxa' (TR1 without historical records) to 100 'taxa' (SUR with all available records). Nonetheless, our phylogenetic models failed to reproduce some diversity features of the data (π and Tajima's D) in data sets in which such diversity was relatively high. Interestingly, the main source of variation in song-element diversity was not the lek identity or the use of historical data, but the alignment strategy (Fig. 5). Phylogenetically informed PRANK-TN93 alignments were generally more compact (Table S1), thus increasing the average number of mismatched positions in song alignments, and thus also increasing π and Tajima's D, but reducing model fit, compared to MAFFT strategies.

In addition to some measures of model adequacy, alignment strategies strongly influenced tree topology (Fig. ?) and age estimates for the ancestor of extant taxa, in the two leks with deeper historical records (Fig. SUR and CCE). In leks with historical records that spanned multiple decades (SUR, CCE, HC1), topologies based on PRANK-TN93 data were markedly distinct from topologies based on phylogenetically naive MAFFT

alignments (Fig ?). In smaller leks with more shallow historical records (BR1 and TR1), the two alignment strategies that enforced a lower substitution rate between vibratory and tonal sounds (PRANK-TN93 and MAFFT-optimal) resulted in relatively similar topologies (Fig ?). In a recent study that compared alternative MSA programs for ancestral protein sequence reconstruction, variations of PRANK and MAFFT algorithms had the highest overall performance (Vialle et al. 2018). However, the two methods were affected in different ways by the underlying substitution process and characteristics of the tree (Vialle et al. 2018). Because leks evolve independently, diverse cultural substitution processes and varying tree topologies are also expected to result in a mixture of alignment performance even within a single species. Our conflicting results between alignment strategies thus underscore a foremost challenge for cultural phylogenetics with behavioral sequences: establishing homology in sequence subunits.

In most phylogenetic analyses, a multiple sequence alignment is treated as an observation, and alignment uncertainty is ignored (REF bali-phy for an exception). However, for phenotypic sequences establishing homology can be a daunting challenge, even when characters are unambiguously coded [Caetano REF]. Furthermore, because culture can change independently of genetic evolution [REF], guiding phenotypic sequence alignments with independently estimated molecular trees might be problematic. Comparing alternative alignment strategies across multiple cultural data sets as we have done here, is a step towards understanding the robustness of cultural phylogenetic analyses. However, as for phylogenetic models, developing tools to assess the adequacy of alignment methods is crucial for a broader application of phylogenetic approaches to cultural sequences. In the case of socially learnt bird songs, alignment methods can be informed by a growing mechanistic understanding of learning, sound production and song function [REFs]. In humans, alignment methods have been tailored to particular cultural domains and used to identify large-scale patterns in melody evolution (Savage et al. 2022) and to resolve deep ancestral relationships between language families (Jäger 2015). In addition to developing specific alignment approaches for different cultural domains, phylogenetic methods that incorporate alignment uncertainty can be implemented to reduce biases in homology hypotheses of behavioural sequences (Caetano and Beaulieu 2020).

Our analyses of socially learnt bird songs make three further assumptions, dictated by the phylogenetic models we sought to apply. First, we discretised song spectrograms of variable duration (X - Xs) into 20 characters, each belonging to one of six categories. This conveniently allowed us to model song evolution as a substitution process, and was justified by the distinctiveness sound categories and high repeatability of character coding (Fig. Supporting text?). However, focusing on substitutions may obscure some **and perhaps important?** cultural changes in song composition. When learning birds modify their song template by extending or contracting particular segments, these changes will introduce indels in the alignment which

do not contribute to the modelled evolutionary process. **We could use a sentence here about how common these changes in song length are, both on LBH and other birds to then say something about how big of an issue this could be.**

Second, we assumed that song elements along a sequence evolve independently of one another. Mechanical, cognitive and functional constraints on song sequence my falsify this assumption. For example, it is likely that if a bird slows down the first trill of a series of consecutive trills, subsequent trills will also be replaced as **biological reason and REF here if there is one**. Similarly, rapid alternations between tonal and vibratory sounds may be strenuous [REF **if this makes any sense**], thus biasing substitution rates based on the identity of nearby sites. The consequences of such site epistasis have been studied mainly in the context of protein evolution, where structure and function generate coevolutionary dynamics among sites [REFs]. In protein evolution, incorrectly assuming site independence can result in biased estimation of site-wise substitution rates [REF], with potential consequences for ancestral sequence reconstruction, clock rate estimation and topological inference. Properties of songs and other behavioral sequences that arise from interactions between subunits may create transmission biases for particular subunit combinations [REF?], potentially resulting in model mispecification of traditional substitution models. Evolutionary models that relax the assumption of site independence may thus also be of applicability for cultural phylogenetics.

Finally, our study assumes that songs diversify in a tree-like manner, with no horizontal transfer. For the Long-billed Hermit system, this means that juvenile males introduce changes on their learning templates, independently of other songs that they may have heard in the same lek. We do not know if males of the Long-billed Hermit indeed copy and potentially modify a sing song template, or if new songs can and are often formed by combining elements from multiple templates. **Maybe a sentence here about what we do know about the learning process.** Nonetheless, it is possible that cultural phylogenetic models are robust to a degree of horizontal transfer, as suggested by the relatively high absolute fit of these models to song data (Table 1). An early study based on cladistics showed that data sets across different cultural domains had a similar fit to a tree-like diversification process as biological data sets [REF Collard]. Further assessments of the robustness of cultural phylogenetics to observed levels of horizontal transfer and reticulate evolution are warranted, as well as the exploration of recently developed methods that can accommodate such processes [REF to that french paper].

Culture can change rapidly compared to morphology [REF]. Thus, cultural phylogenetics studies can often tap on a relatively rich historical record, obtained from archeological preservation (in humans) or from longitudinal studies. Historical records in our longitudinal sampling of Long-billed Hermit songs have some advantages over traditional fossils, such as complete character sequences and no stratigraphic uncertainty. Our study also

shares some of the challenges of phylogenetic inference with fossils, like time-heterogeneous preservation. We thus expected that historical song records would impact estimates of node ages and diversification rates to at least the same extent as total-evidence approaches under the FBDP influence the same parameters in organic lineages. We found that incorporating historical data resulted in more precise root age estimates, as seen across total-evidence dating analyses under the FBDP [REFs to Total evidence]. Furthermore, by sampling ancestral lineages that formed and went extinct before the present, the historical song record also accelerated estimates of both speciation and extinction, but without effects on net diversification (Fig. 7). Our results therefore suggest that unlike organic evolution, which typically follows a trend of increasing diversity over time, social factors, such as sexual selection and kin recognition, may impose net-zero diversification in stable cultural systems [REF?].

We were also interested in understanding how decisions on how to use available historical records would impact estimation of node ages. We found that including older but sporadically sampled records resulted in older estimates of crown and root ages for all leks with such data (CCE, HC1 and SUR; Fig. 7). In contrast, accounting for all historical occurrences, including identical sequences sampled on different years, increased only the precision of age estimates (crown and origin) in the two largest leks (CCE and SUR; Fig. 9). In fact, the main consequence of using all historical occurrences was to increase support for a relaxed clock model (Fig. 3), as clock rates must differ between branches separating identical songs and branches in which substitutions occurred.

Simulation studies show that accounting for all possible fossil occurrences [REF OReily] and sampling fossils at regular time intervals [REF 2020Sim] bolster the accuracy of phylogenetic inference of organic evolution under the FBDP. These sampling practices, especially the latter, may be prohibitive for some organic clades. Here, we based our analyses on pre-existing data, most of which was collected before the methods we used were even developed. However, future cultural phylogenetics studies based on longitudinal data can incorporate these best sampling strategies in their design, and more systematically assess how the availability and regularity of historical records impact age estimates of cultural lineages.

Conclusions

Acknowledgements

...B.W. is funded by an International Postdoc Grant (2019-06444) from the Swedish Research Council (Vetenskaprådet)... UCR (VI) project...

Author contributions

B.W. and **M.S.A.** conceived the study. **M.S.A.** collected the data with support from **A.R.G.**. **M.S.A.** and **B.W.** analyzed the data and wrote the manuscript with contributions from **A.R.G.**.

References

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

Araya-Salas M., Odom K. 2022. PhenotypeSpace: An r package to quantify and compare phenotypic trait spaces.

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.

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.

Carpenter B., Gelman A., Hoffman M.D., Lee D., Goodrich B., Betancourt M., Brubaker M., Guo J., Li P., Riddell A. 2017. Stan: A probabilistic programming language. Journal of Statistical Software. 76.

Carstens B.C., Smith M.L., Duckett D.J., Fonseca E.M., Thomé M.T.C. 2022. Assessing model adequacy leads to more robust phylogeographic inference. Trends in Ecology & Evolution.

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.

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.

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

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

556 Höhna S., Landis M.J., Heath T.A., Boussau B., Lartillot N., Moore B.R., Huelsenbeck J.P., Ronquist F. 2016.
557 RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification
558 language. Systematic Biology. 65:726–736.

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

561 Jäger G. 2015. Support for linguistic macrofamilies from weighted sequence alignment. Proceedings of the
562 National Academy of Sciences. 112:12752–12757.

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

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

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

570 Kershenbaum A., Blumstein D.T., Roch M.A., Akçay C., Backus G., Bee M.A., Bohn K., Cao Y., Carter G.,
571 Cäsar C., Coen M., Deruiter S.L., Doyle L., Edelman S., Ferrer-i-Cancho R., Freeberg T.M., Garland
572 E.C., Gustison M., Harley H.E., Huetz C., Hughes M., Hyland Bruno J., Ilany A., Jin D.Z., Johnson
573 M., Ju C., Karnowski J., Lohr B., Manser M.B., Mccowan B., Mercado E., Narins P.M., Piel A., Rice
574 M., Salmi R., Sasahara K., Sayigh L., Shiu Y., Taylor C., Vallejo E.E., Waller S., Zamora-Gutierrez V.,
575 Akçay Ç., Backus G., Bee M.A., Bohn K., Cao Y., Carter G., Cäsar C., Coen M., Deruiter S.L., Doyle
576 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.

Khouri Z., Gillung J.P., Kimsey L.S. 2022. The evolutionary history of mammoth wasps (Hymenoptera: Scoliidae). BioRxiv.

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.

May M.R., Contreras D.L., Sundue M.A., Nagalingum N.S., Looy C.V., Rothfels C.J. 2021. Inferring the total-evidence timescale of marattialean fern evolution in the face of model sensitivity. Systematic Biology. 70:1232–1255.

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.

Muff S., Nilsen E.B., O'Hara R.B., Nater C.R. 2021. Rewriting results sections in the language of evidence. Trends in Ecology & Evolution.

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

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

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

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

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

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

635 Richards E.J., Brown J.M., Barley A.J., Chong R.A., Thomson R.C. 2018. Variation across mitochondrial
636 gene trees provides evidence for systematic error: How much gene tree variation is biological? Systematic
637 Biology. 67:847–860.

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

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

642 Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Mathematical Biosciences. 53:131–147.

643 Sagart L., Jacques G., Lai Y., Ryder R.J., Thouzeau V., Greenhill S.J., List J.-M. 2019. Dated language
644 phylogenies shed light on the ancestry of Sino-Tibetan. Proceedings of the National Academy of Sciences.
645 116:10317–10322.

646 Savage P.E., Passmore S., Chiba G., Currie T.E., Suzuki H., Atkinson Q.D. 2022. Sequence alignment of folk
647 song melodies reveals cross-cultural regularities of musical evolution. Current Biology. 32:1395–1402.

Schliep K.P. 2011. Phangorn: Phylogenetic analysis in R. Bioinformatics. 27:592–593.

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

Stiles F.G., Wolf L.L. 1979. Ecology and evolution of lek mating behavior in the long-tailed hermit hummingbird. Ornithological Monographs. 27.

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

Ten Cate C. 2021. Re-evaluating vocal production learning in non-oscine birds. Philosophical Transactions of the Royal Society B. 376:20200249.

Vialle R.A., Tamuri A.U., Goldman N. 2018. Alignment modulates ancestral sequence reconstruction accuracy. Molecular Biology and Evolution. 35:1783–1797.

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.

Wickham H. 2016. ggplot2: Elegant graphics for data analysis. Springer-Verlag New York.

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.

669 Yu G., Smith D., Zhu H., Guan Y., Lam T.T.-Y. 2017. ggtree: An R package for visualization and annotation
670 of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution.
671 1:28–36.

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

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

Supporting Material

Figure S1. Examples of songs and coding?

Supporting text 1

How do we arrive to 20 segments

Supporting text 2

Repeatability of sound classification

Figure S2. Main repeatability result.

Table S1. MCMC diganostics, add alignment legnth and no. of sequences to the table we have

Figures S3-S32. All model reliability results

Figures S33-S35. Histograms of age and diversification parameters per record type for all alignment strategies.

Figures S36-S38. Histograms of age and diversification parameters per sampling strategy for all alignment strategies.