Supplementary material for:

Comparative analyses of phenotypic sequences using phylogenetic trees

Daniel S. Caetano* and Jeremy M. Beaulieu

*contact: caetanods1@gmail.com

List of figures:

List of tables:

Table of contents:

1) Simulation results - Figures and tables

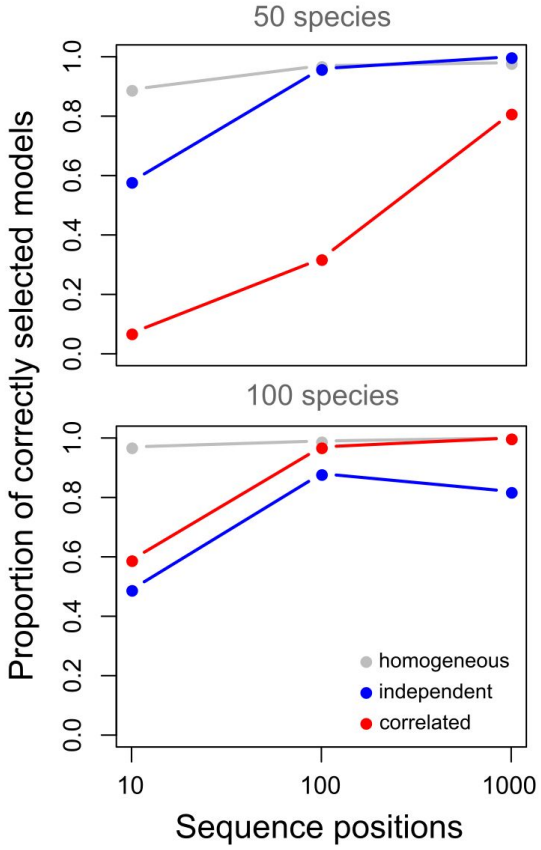

Figure S1: Results from 100 simulation replicates under each of the generating models. Gray lines show results for the homogeneous rate model. Blue lines show results for the independent rates (Gamma) model. Red lines show results for the correlated model. Top plot shows results with 50 species and bottom plot with 100 species. The proportion of correctly selected models was computed by counting the number of times, across the 100 replicates, that the generating model showed $\Delta AIC < 2$.

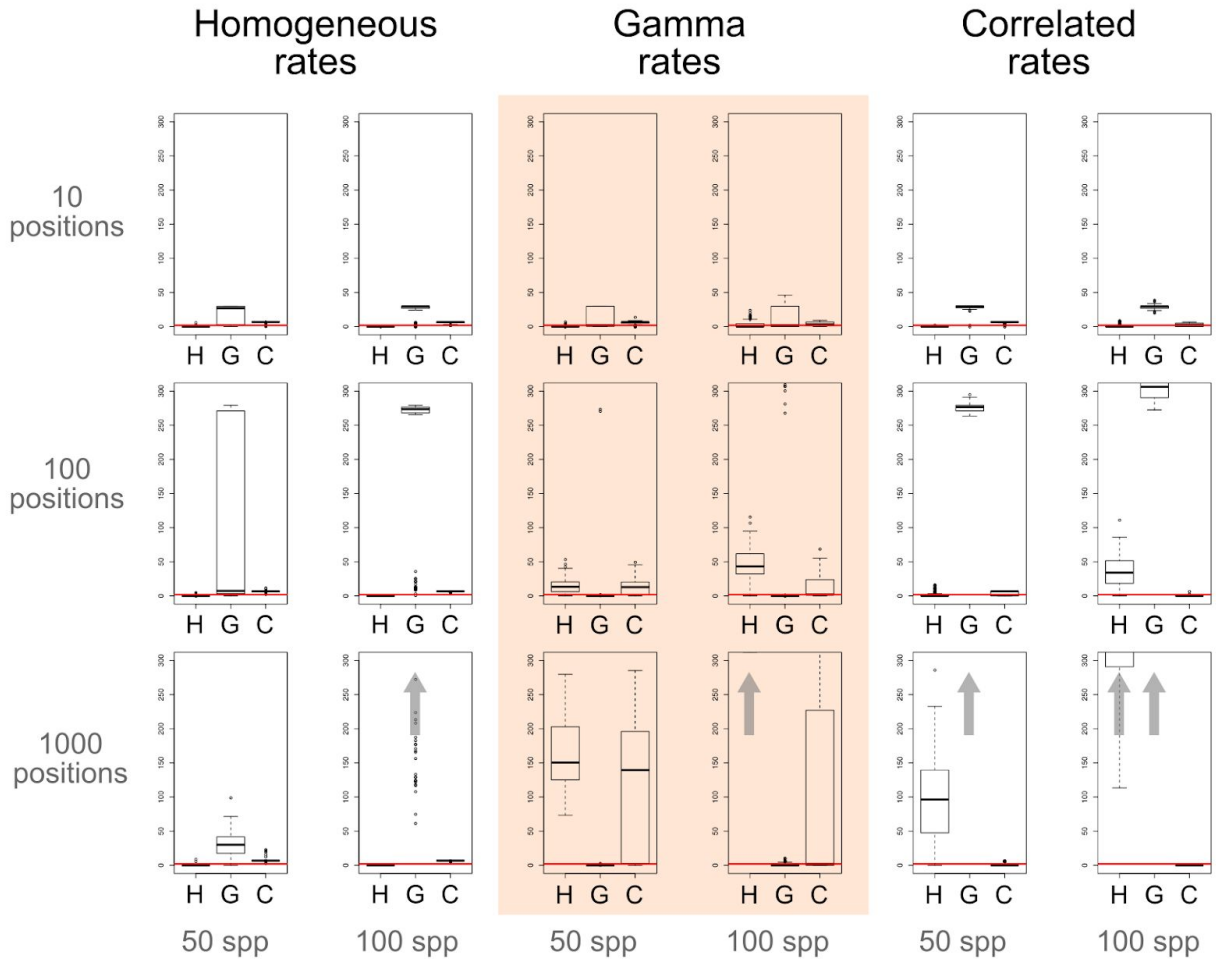

Figure S2: Distribution of pairwise ΔAIC for each of the models across simulation treatments. Each plot show results for a window from 0 to 300 ΔAIC units. Any result with mean value above 300 ΔAIC units was omitted from the plot and represented by a grey arrow. The salmon square in the middle highlights the results for the independent rates (Gamma) model. To the left are results for the homogeneous model and to the right the results for the correlated rates model. Top row shows results for simulations with only 10 positions, middle row with 100 positions and bottom row with 1000 positions. Each column separates the simulations between phylogenetic trees with 50 and 100 species. The red lines mark the $\Delta AIC < 2$ threshold which groups the models favored by the data.

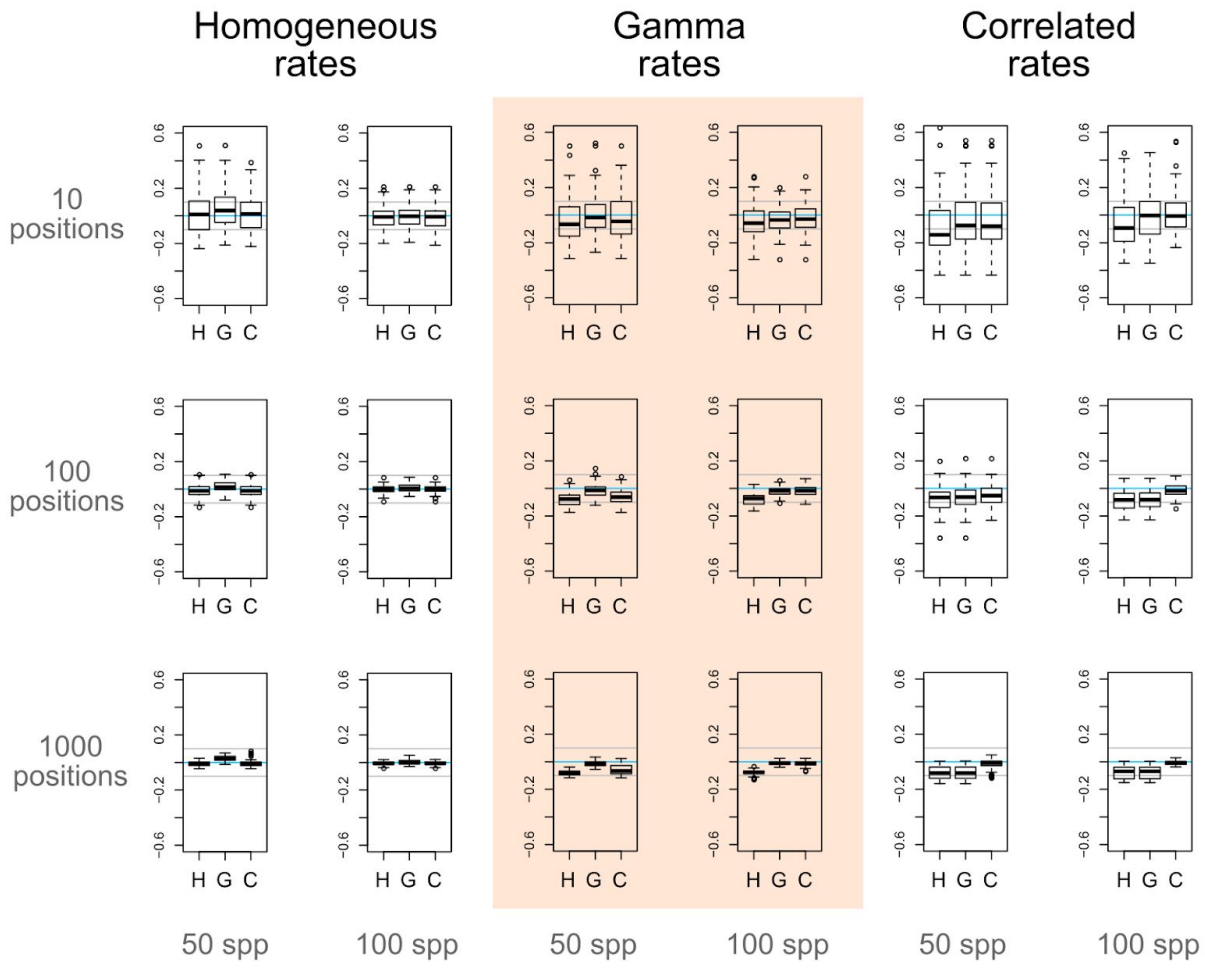

Figure S3: Distribution of parameter estimates for each of the models across simulation treatments showed as the relative average distance (computed as the proportion of the true parameter) from the true rates of evolution for each sequence position. Light blue lines mark the 0 point (i.e., no difference in average rates). Grey lines mark the 10% deviance limit above or below the true rates. The salmon square highlights the results for the independent rates (Gamma) model. To the left are results for the homogeneous model and to the right the results for the correlated rates model. Top row shows results for simulations with only 10 positions, middle row with 100 positions and bottom row with 1000 positions. Each column separates the simulations between phylogenetic trees with 50 and 100 species.

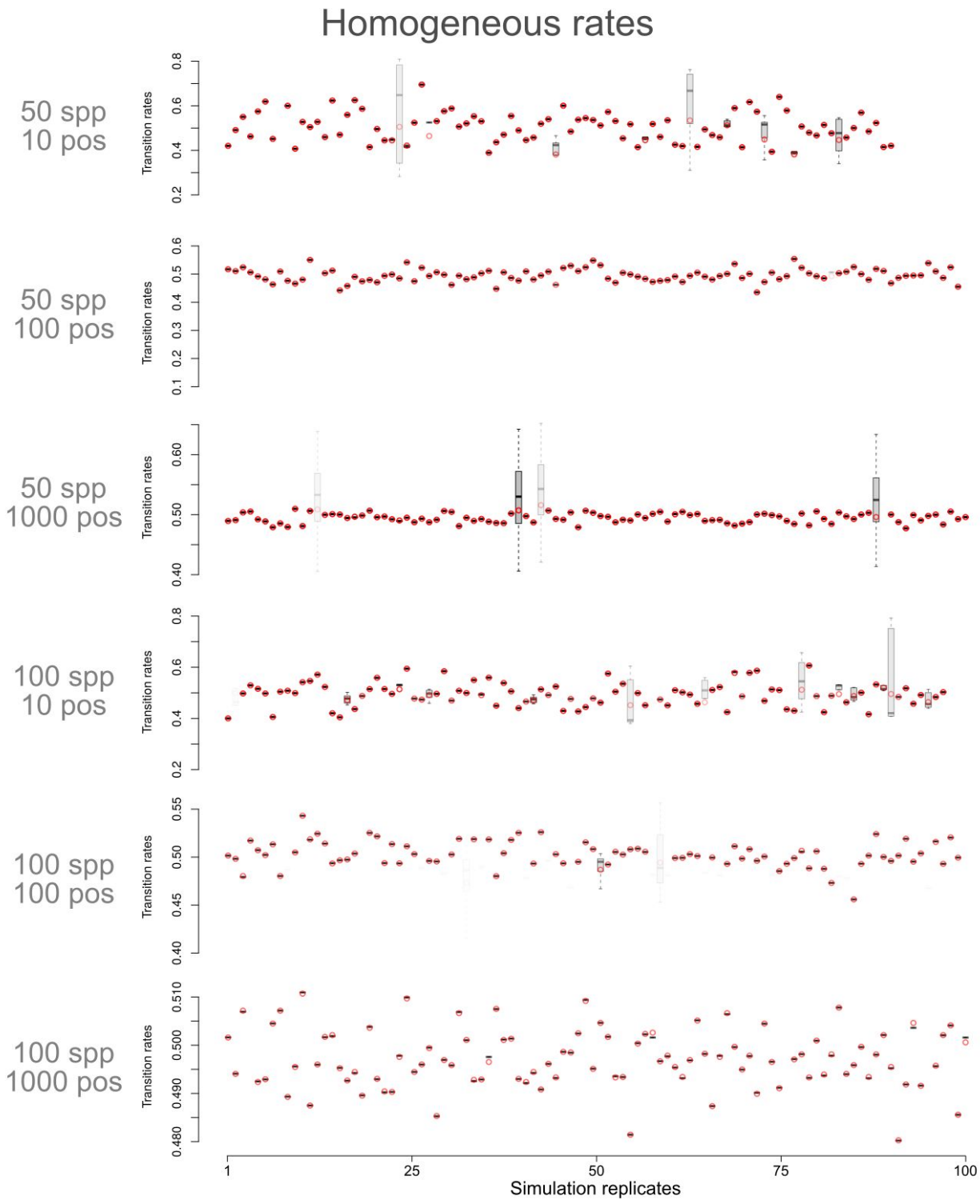

Figure S4: (Caption on next page)

Figure S4: Distribution of rate estimates for the single rate model (red circles) and correlated rates model (overlaid box plots) for datasets simulated under a single rate model. Each plot shows results for a different simulation treatment varying the number of species (spp) and the number of sequence positions (pos). The x-axis of the plots show results for the 100 simulations replicates for each of the treatments (order is arbitrary). Red circles are transition rates under the single rate model for each simulation. Rates under the correlated model are shown as box plots and black horizontal lines represent box plots with no variation (i.e., cases in which the correlated model recovered a single rate across the sequence). The transparency of the color for each simulation point (red circle + box plot pair) is proportional to the ΔAIC between models. In other words, opaque symbols show models with similar AIC and transparent symbols represent models with distinct AIC scores.

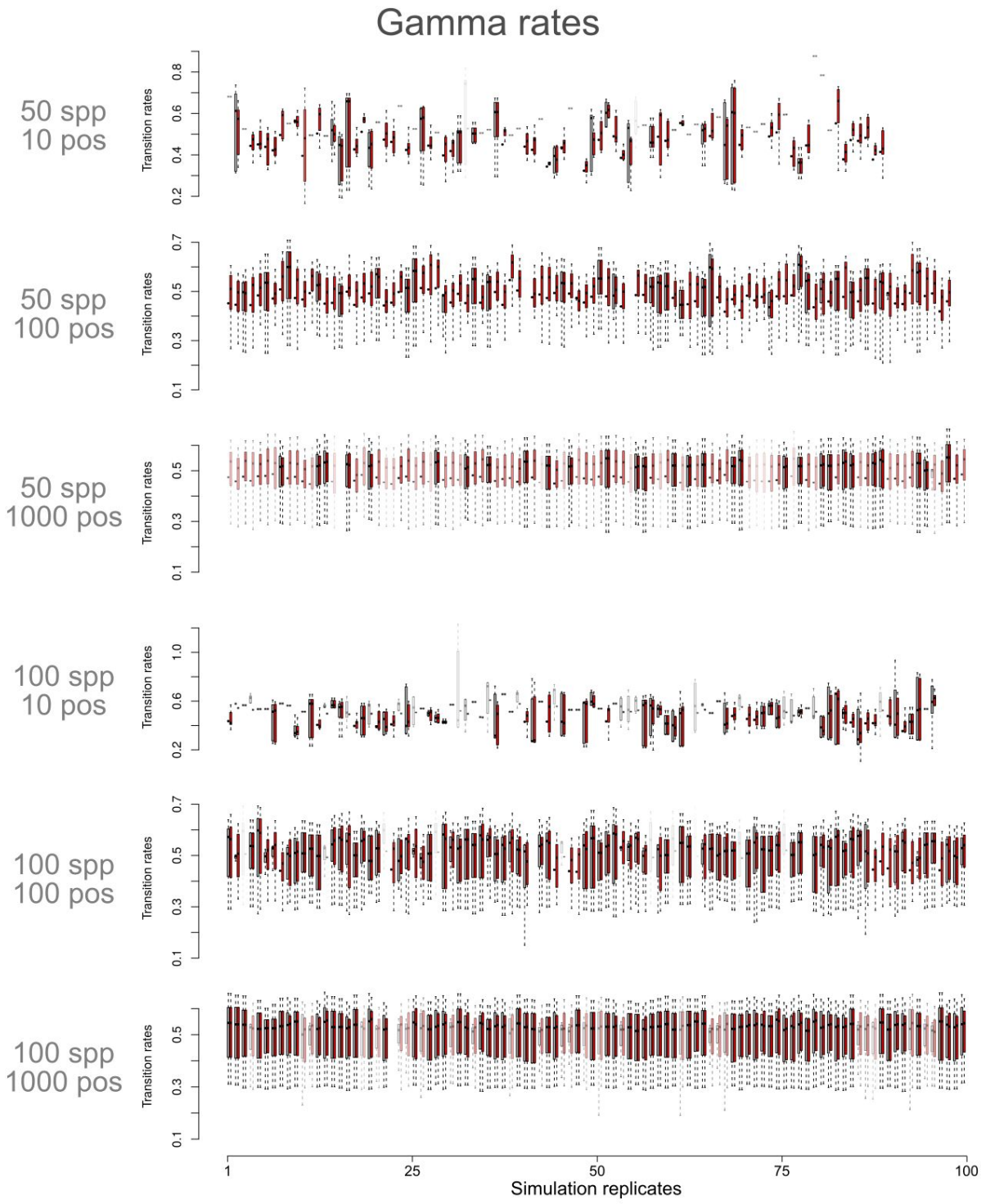

Figure S5: (Caption on next page)

Figure S5: Distribution of rate estimates for the gamma model (red box plots) and correlated model (grey box plots) for datasets simulated with rate heterogeneity along sequence positions (i.e., with Gamma distributed rate categories). Each plot shows results for a different simulation treatment varying the number of species (spp) and the number of sequence positions (pos). The x-axis of the plots show results for the 100 simulations replicates for each of the treatments (order is arbitrary). Each pair of red and grey box plots represents the distribution of rates under the gamma and correlated model, respectively. Each pair of box plots (denoting results from the same simulation replicate) is separated by a small white space. The transparency of the color for each simulation (red + grey box plot pair) is proportional to the ΔAIC between models. In other words, opaque symbols show models with similar AIC and transparent symbols represent models with distinct AIC scores.

2) Empirical case: Evolution of calling song in Gryllus crickets

2.1) Estimating the Gryllus phylogeny from molecular data

We downloaded sequences for mitochondrial 16S and cytochrome B (cytb) from GenBank following accession numbers reported by Robillard et al. (2006 - see their Table 1). We aligned the 16S and cytb sequences using MAFFT (Katoh et al. 2013) and fitted molecular evolution models for each gene using jModelTest 2 (Darriba et al. 2012). Results showed support for the GTR+I+G model for both genes under AIC and BIC model selection criteria.

In order to infer a tree with ultrametric branch lengths and incorporate uncertainty in tree estimation we used BEAST2 (Bouckaert et al. 2014) with a birth and death tree prior and relative branching times. Our initial analyses using default priors showed that transition rates between C and G as well as G and T were near 0. This creates issues with the mixing of the chain because the Gamma distribution does not allow rates to be 0. In other words, the Markov-chain Monte Carlo (MCMC) simulation starts to sample too close to the hard bound of the prior distribution. For instance, the initial MCMC returned an Effective Sample Size (ESS) of less than 10 units for these transition rates.

We updated our selection of priors on the transition rates to a uniform distribution with range informed by the results from our preliminary analysis with the same data (see Figure S6). Initial values for the MCMC chain were set to the mean of the posterior distribution for each of the parameters estimated using the initial default priors. The range of the improper uniform prior distributions were set to be proportional to the range of the posterior distribution under BEAST2 default priors. Then, we ran the MCMC for 10 Million generations and excluded 25% of the samples as burn-in. We checked the trace of the posterior, likelihood, and priors to check for convergence of the chain as well as the ESS of each parameter of the model to check for good mixing. We also computed the ESS for the tree topology using the R package “rwty” (Warren 2017).

The resulting phylogenetic tree showed good posterior support, with the majority of nodes showing posterior probabilities above 0.9 (Figure S7). When compared with the trees reported by Robillard et al. (2006), the topology of the Maximum Clade Credibility (MCC) tree is closer to the phylogeny using molecular data and acoustic characters (their figure Figure 2A). The maximum clade credibility tree reported on Figure S7 is congruent with results from Huang et al. (2000).

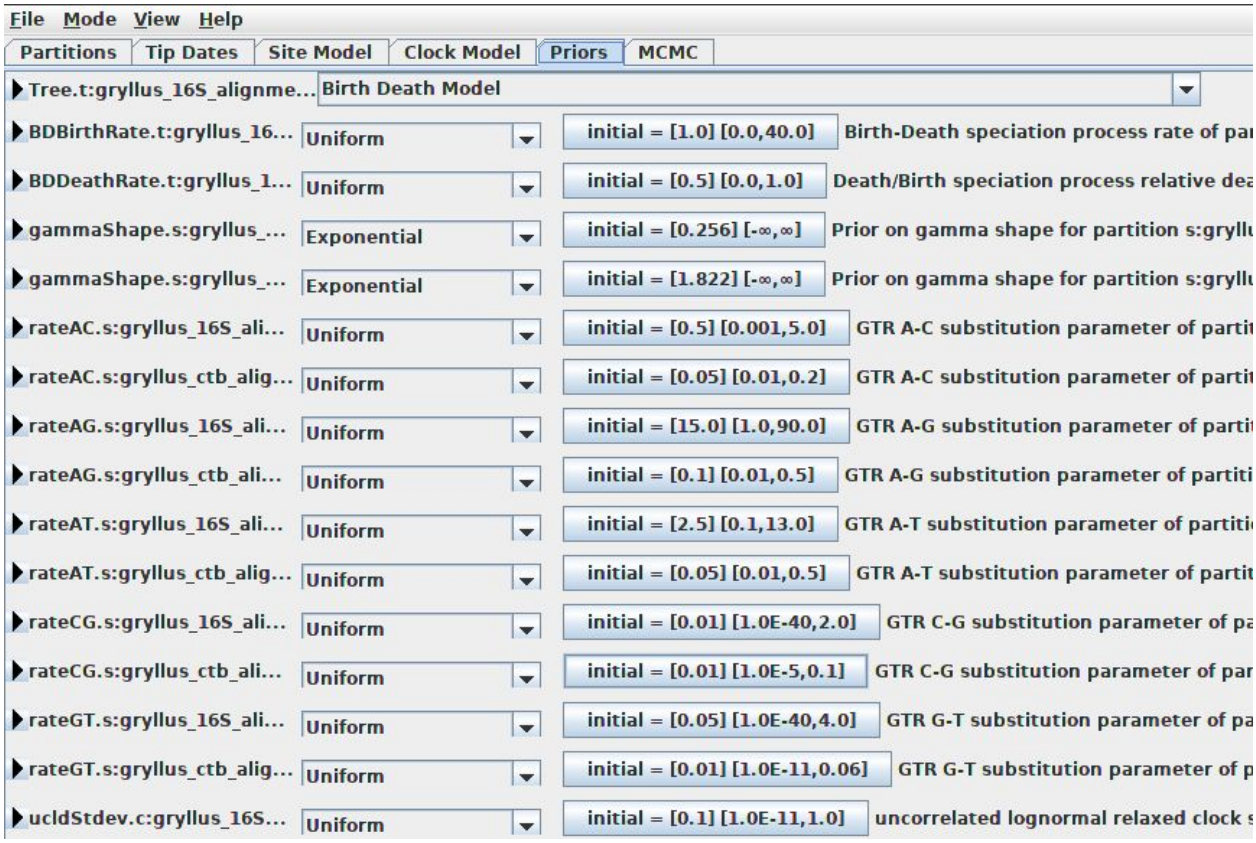

Figure S6: Starting values and prior distributions for the parameters of the BEAST2 analyses using 16S and cytb sequences for *Gryllus* crickets. Figure is a screen capture from the BEAUti user interface set to the "Priors" tab.

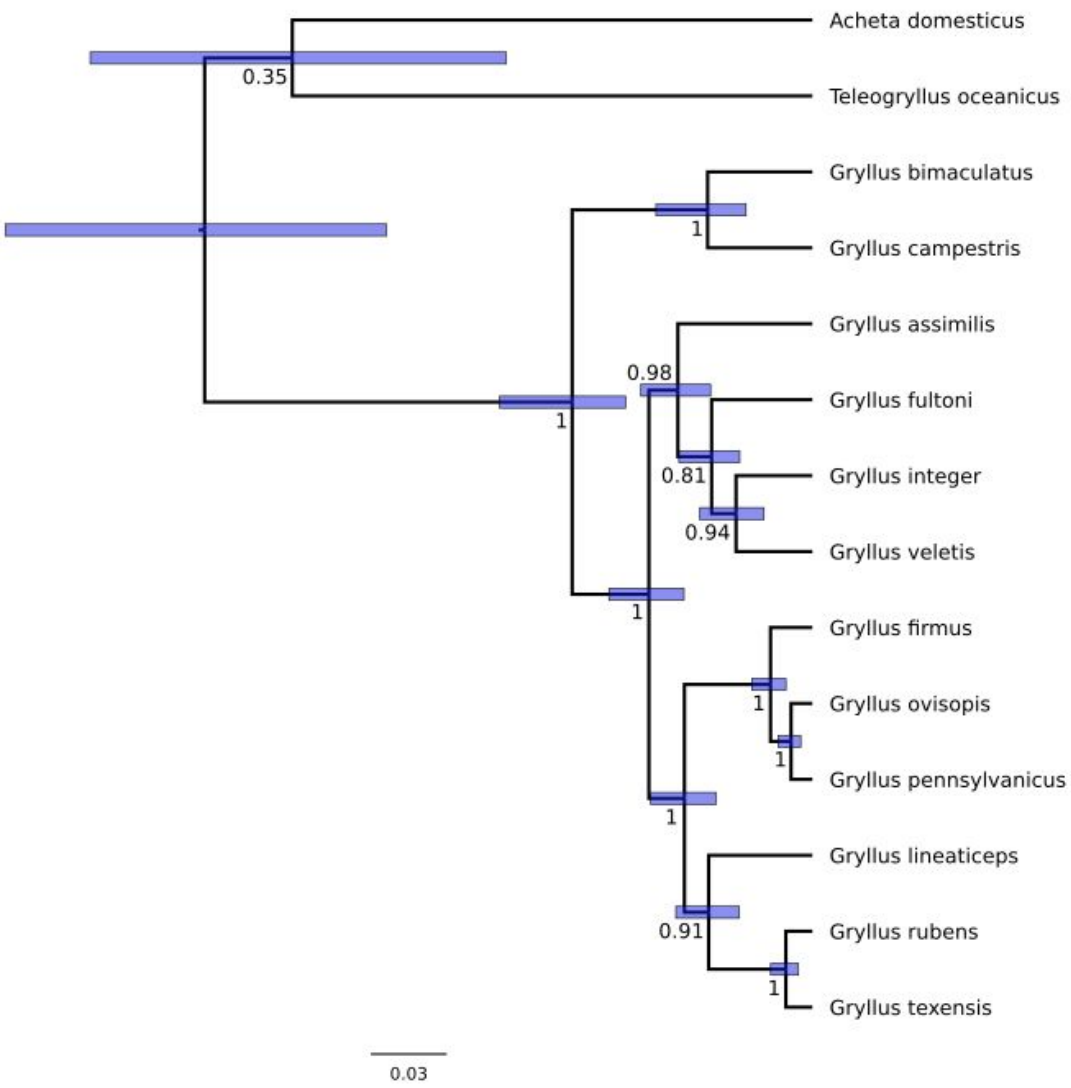

Figure S7: Maximum Clade Credibility (MCC) tree. Numbers at the nodes show posterior node support computed as the proportion of times each node was observed across the posterior samples. Purple bars show the 95% credible interval for the node heights in units of relative branching times.

2.2) Rescaling cricket songs sequences to units of time

Robillard and colleagues (2006) coded the temporal pattern of the male *Gryllus* calling songs using a single character state (coded as A) for each syllable, or pulse. Then they implemented two approaches to code the silent regions. First they used a single character (coded as G) to denote a silent region independent of its length (i.e., DO1 coding). Alternatively, the authors used a string of silent characters (i.e., a string of G) to code the silent segment with a number of characters equal to the time interval divided by the syllable period (i.e., DO2 coding). Here we used only the DO2 character coding in our analyses because, as discussed by Robillard et al. (2006), the DO2 code scheme takes into account the time information of the interchirp regions and performs better than the DO1 coding scheme. We refer to the DO2 coding as "original sequences" in the main text.

Each song sequence was recorded for an interval of 5 seconds. Under the DO2 coding, the sequence for a species has a number of characters equal to 5 seconds divided by the time period of a single pulse. In other words, the sequences reported by Robillard et al. (2006) are in units of pulse. Since the time period of a pulse varies between species, the unaligned sequence lengths also vary. Here we explored the temporal pattern of the male calling songs by transforming the sequences from units of pulse back to units of time (seconds). For this we computed the time period for the pulse of each species and rescaled the sequences such that each sequence would have the same length. Thus, in the time rescaled data reported here, a sound character is not equivalent to a pulse, but it is a representation of a common unit of time between the species. We report results with both types of sequence for each analysis conducted in this study.

2.3) Selecting song bout sequences

The calling song of male *Gryllus* crickets is a complex trait emerging from repeated pulses separated by periods of silence. A pulse is a complete movement of the wings, which is the homologous behavior that produces the sound (Otte 1992, Robillard et al. 2006). When pulses are repeated in a tight sequence, with no intervals (or intervals of very short length) a chirp (or trill) is produced (see Otte 1992). The silent interval between chirps (or trills) is called interchirp silence. Here we define a song bout as a higher level repeated pattern for which a group of chirps (or trills) and interchirp silences constitute the units of repetition. Figure 1 shows the whole 5 seconds of song recordings for each of the species of *Gryllus* included in this study. The majority of species shows a simple song bout, comprised by a single chirp (or trill) separated by an interchirp silence region which repeats multiple times through the recorded sequence. Species *G. campestris*, *G. bimaculatus*, *G. texensis*, *G. lineaticeps*, *G. firmus*, *G. veletis*, and *G.*

assimilis belong to this group. In contrast, *G. integer* and *G. fultoni* have more complex song bouts. Figure 1 shows that, for both species, there is a region of short chirps connected by short interchirp silences which is flanked by much longer interchirp silences. We defined these regions as the song bouts. Finally, it is not possible to evaluate the song bout pattern of *G. rubens* because it's longer song bout was not fully recorded within the 5 seconds window.

We manually separated each song bout from the complete song sequences for all species. This generated a file with multiple song bouts per species, except for *G. integer* and *G. fultoni* whose sequences only presented a single bout. Then, in order to include within sequence variance in song bouts, we produced data sets by randomly drawing one song bout per species. We were able to produce 100 unique data sets by mixing song bouts of all species.

2.4) Alignment of cricket songs using MAFFT

The default algorithm in MAFFT (Katoh et al. 2013) produces an alignment based on a guide tree, which is inferred from the sequences as part of the algorithm using different approaches. MAFFT is a distance-based alignment approach that, by default, applies the standard BLOSUM62 score matrix (Henikoff and Henikoff 1992). In contrast, there is no standard cost matrix for the alignment of non-molecular sequences.

The *Gryllus* male calling song sequences used in the present study were coded with four states (following Robillard et al. 2006): sound characters (A), interchirp silence (G), and intrachirp silence (C). Robillard and colleagues (2006) also used the states R and M to indicate polymorphic states: either sound or intrachirp silence (R) and either sound or interchirp silence (M). In order to align these sequences we used the option `` `--textmatrix` `` implemented in MAFFT (supported on versions ≥ 7.120 - Katoh et al. 2013). This option allows the user to define a custom scoring matrix, which substitutes the default score matrix. In this matrix, scores > 1.0 indicate an important pairing and scores between 0.0 and 1.0 are set for different elements that share similarities (with larger values carrying more weight than lower values). Finally, mismatches between important and dissimilar elements are set to negative values. With this scoring scheme it is possible to use the alignment algorithm in MAFFT without the need to assume patterns associated with molecular evolution processes.

Table S2 shows the custom scores used in this study. We set cost penalties such that matches between sound elements would be the most informative, followed by matches between sound elements and the R states (either sound or intrachirp silence). Then, we set matches between either intra- or interchirp silences as slightly less informative than sound elements. Further combinations were set following these same general rules.

We repeated the alignment procedure for the complete song sequences with 100 phylogenetic trees randomly sampled from the posterior distribution of trees inferred using MCMC in order to incorporate the effect of uncertainty in topology and branch lengths. In the case of the song bouts, we repeated the alignment across 100 data sets incorporating variance on the song bout expression across the recorded sequences and conditioned on the MCC tree. The alignment with the original sequences showed a distribution of sequence length varying between 1151 and 1245 positions (Figure S8A). In contrast, the alignment using the time rescaled sequences based on the same sample of trees had a minimum of 1501 and maximum of 1717 positions (Figure S8B). Finally, the alignment using the song bout sequences varied between 261 and 290 positions for the original sequences (Figure S9A) and 399 and 442 positions for the time-rescaled sequences (Figure S9B).

Table S1: Custom score matrix used to perform the alignment of *Gryllus* song sequences in MAFFT. Column 'Pair code' shows the HEX code required by MAFFT to identify character symbols. Column 'Pair states' shows the correspondent pair of states. Finally, column 'Score' shows the relative score for matches and mismatches.

Pair code	Pair states	Score
0x41 0x41	A A	1.5
0x41 0x52	A R	1.3
0x52 0x41	R A	1.3
0x47 0x47	G G	1
0x43 0x43	C C	1
0x47 0x52	G R	0.7
0x52 0x47	R G	0.7
0x43 0x52	C R	0.6
0x52 0x43	R C	0.6
0x43 0x47	C G	0
0x47 0x43	G C	0
0x41 0x47	A G	- 0.5
0x47 0x41	G A	- 0.5
0x41 0x43	A C	- 0.5
0x43 0x41	C A	- 0.5

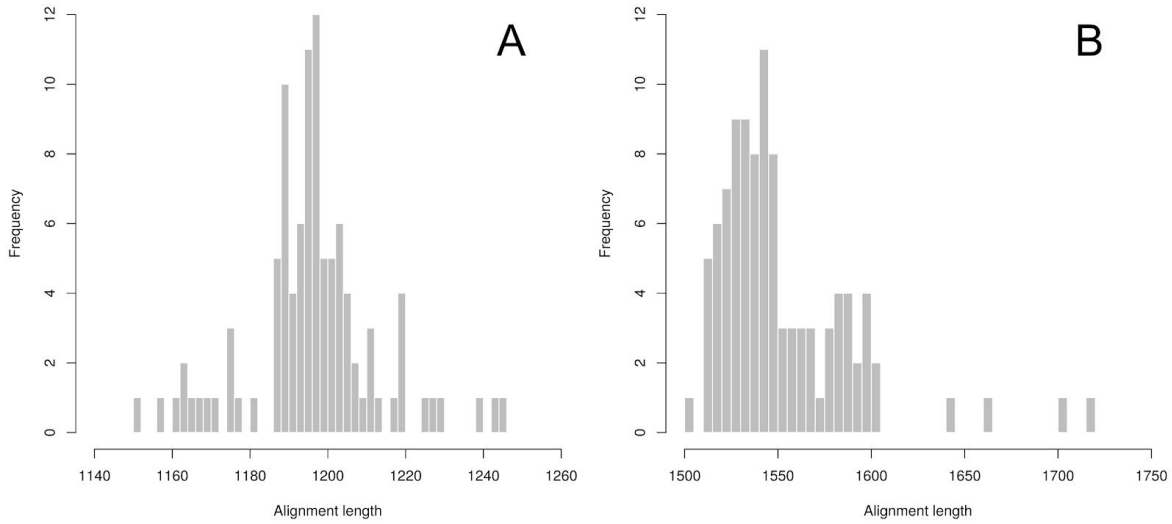

Figure S8: Distribution of alignment lengths for complete song sequences using MAFFT over a sample of 100 trees from the posterior distribution. A) Original sequences (DO2). B) Time rescaled sequences.

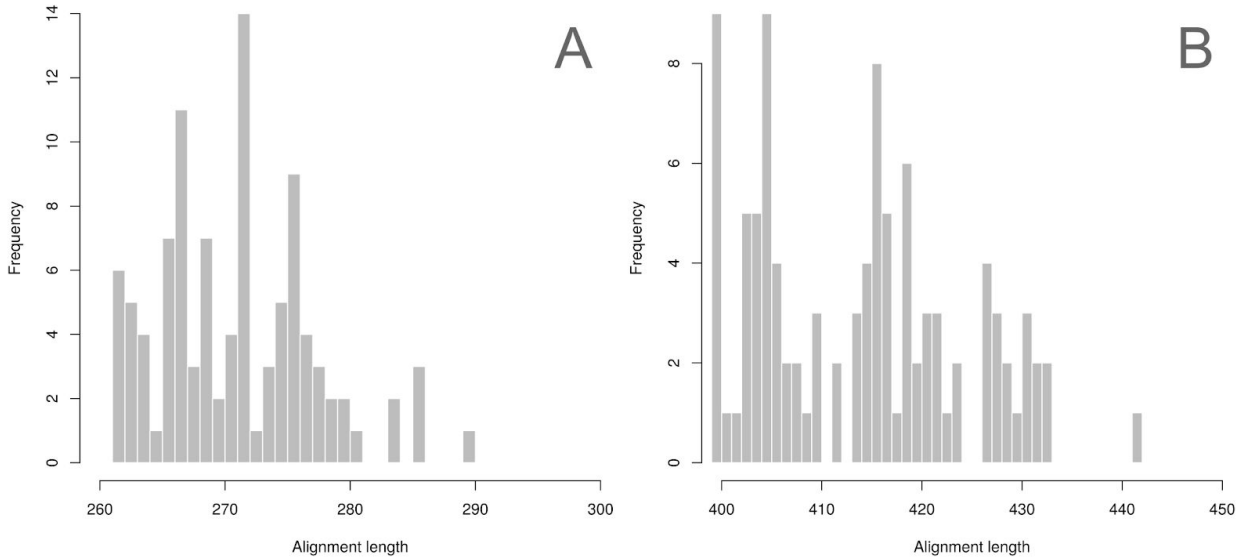

Figure S9: Distribution of alignment lengths for the song bout data. A sample of 100 data sets was produced by random sub-sample of song bouts from the complete song sequences. Distance-based alignments were conducted using MAFFT conditioned on the MCC tree. A) Bouts from original sequences (DO2). B) Bouts from time-rescaled sequences.

2.5) Statistical alignment of cricket songs using BAli-Phy

We also performed alignment estimates using BAli-Phy in order to incorporate alignment uncertainty based on a statistical alignment approach conditioned on the Maximum Clade Credibility (MCC) tree. For this we constrained the topology to be the same as the MCC tree, but allowed branch lengths to vary because BAli-Phy estimates branch lengths following the parameters of the model of sequence evolution (Redelings and Suchard 2005). For this we set the following options:

```
# Disable moves on the topology
disable = topology
# Set the phylogeny
tree = Gryllus_only_tree_MCC.tre
```

We ran two independent MCMC chains for each data set (DO2 and time rescaled) and evaluated convergence by checking trace plots of prior and posterior probabilities on Tracer (Rambaut 2018). We also checked if the effective sample size for the parameters were reasonably high (> 500). We computed the Gelman and Rubin's potential scale reduction factor (Gelman and Rubin 1992) for all parameters as implemented in the function `gelman.diag` of the R package "coda" (version 0.19.2, Plummer 2006) for each parameter of the model.

The alignment with the original sequences (Robillard et al. 2006), using BAli-Phy and conditioned on the topology of the MCC tree (Figure S10) showed length varying between 1409 and 1727 positions (Figure S10A). In contrast, the alignment using the time rescaled sequences resulted in a posterior distribution of alignments much longer than the one based on the original sequences (minimum of 2145 and maximum of 2445 positions) (Figure S10B). Finally, the distribution of alignments using the song bout data conditioned on the topology of the MCC tree resulted in alignments between 332 and 514 positions for the original sequences and 500 and 696 positions for the time-rescaled sequences (Figure S11).

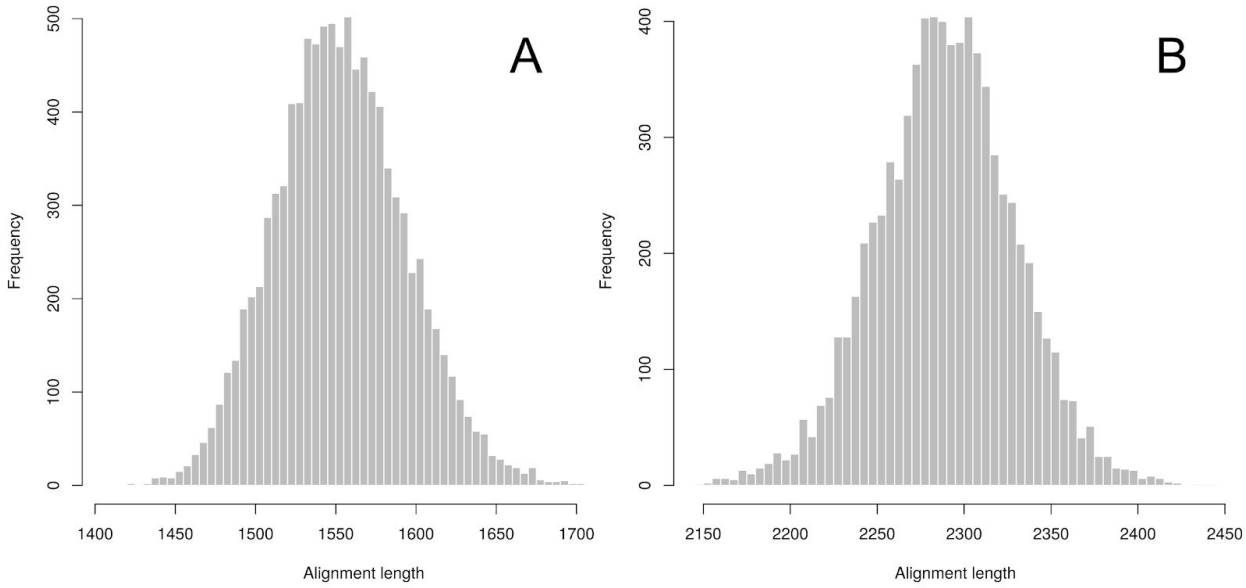

Figure S10: Distribution of alignment lengths for the complete song sequences from the posterior distribution of BAli-Phy conditioned on the topology of the MCC tree. A) Original sequences (DO2). B) Time rescaled sequences.

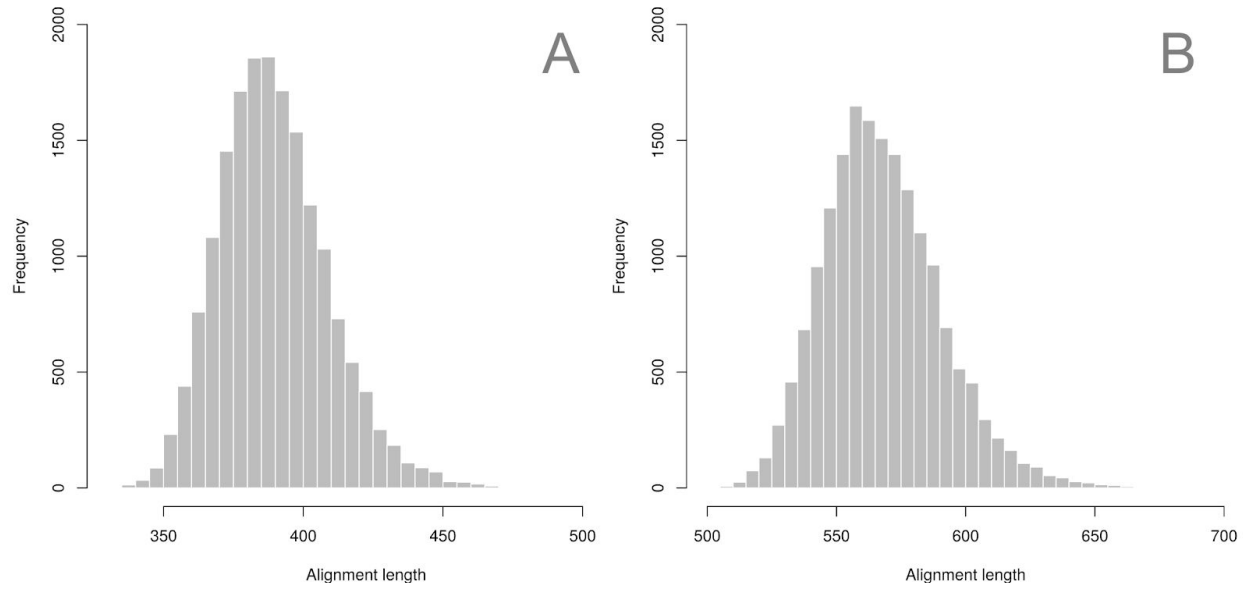

Figure S11: Posterior distribution of alignment lengths for the song bout data from BAli-Phy. A) Bouts from original sequences (DO2). B) Bouts from time-rescaled sequences.

2.6) Support for alternative models

The plot below lists the delta AIC values for the fitted models across all data transformations and alignment approaches implemented on this study. These complement results reported in Table 1 (main text).

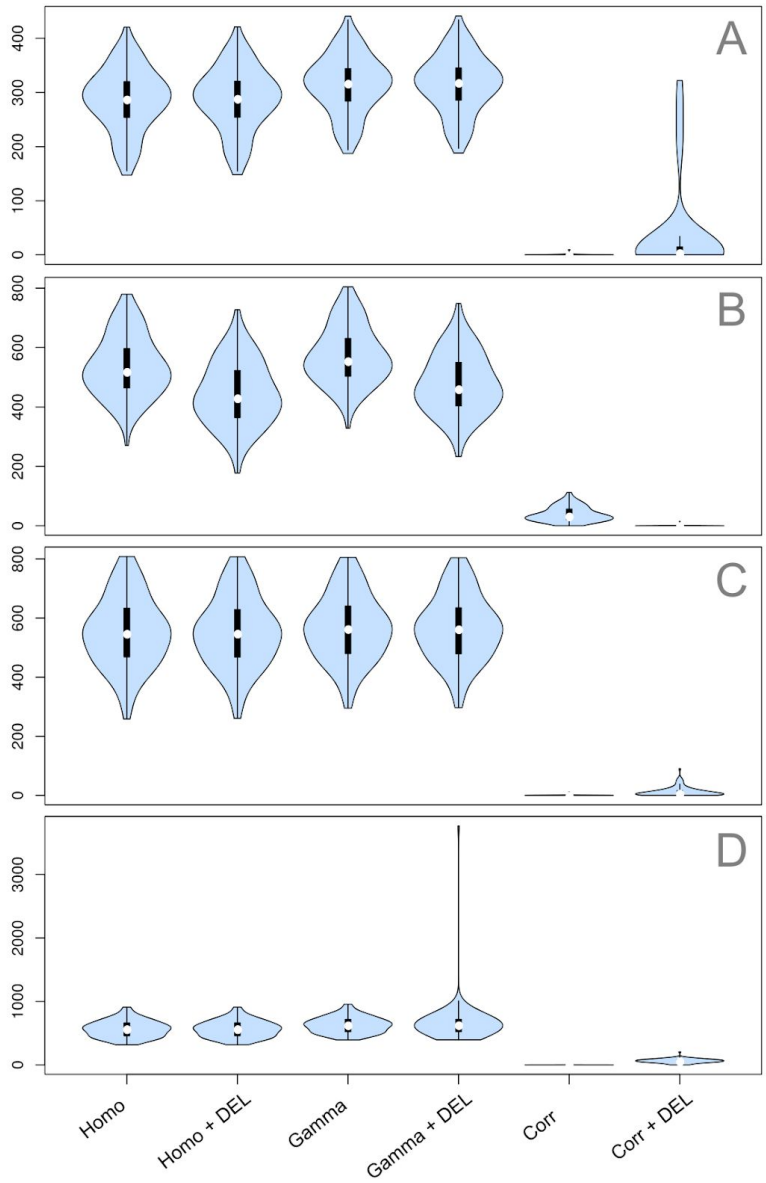

Figure S12: Distribution of delta Akaike Information Criterion (ΔAIC) for models of sequence evolution applied to both original (DO2) and time rescaled song sequences. A) Original sequences and distance-based alignments. B) Rescaled sequences and distance-based alignments. C) Original sequences and model-based alignments. D) Rescaled sequences and model-based alignments. Distributions on A and B were estimated across a random sample of 100 trees. Distributions on C and D were estimated across a random sample of 100 alignments from the posterior distribution of alignments conditioned on the topology of the maximum clade credibility tree (MCC). See Table 1 for a description of the models.

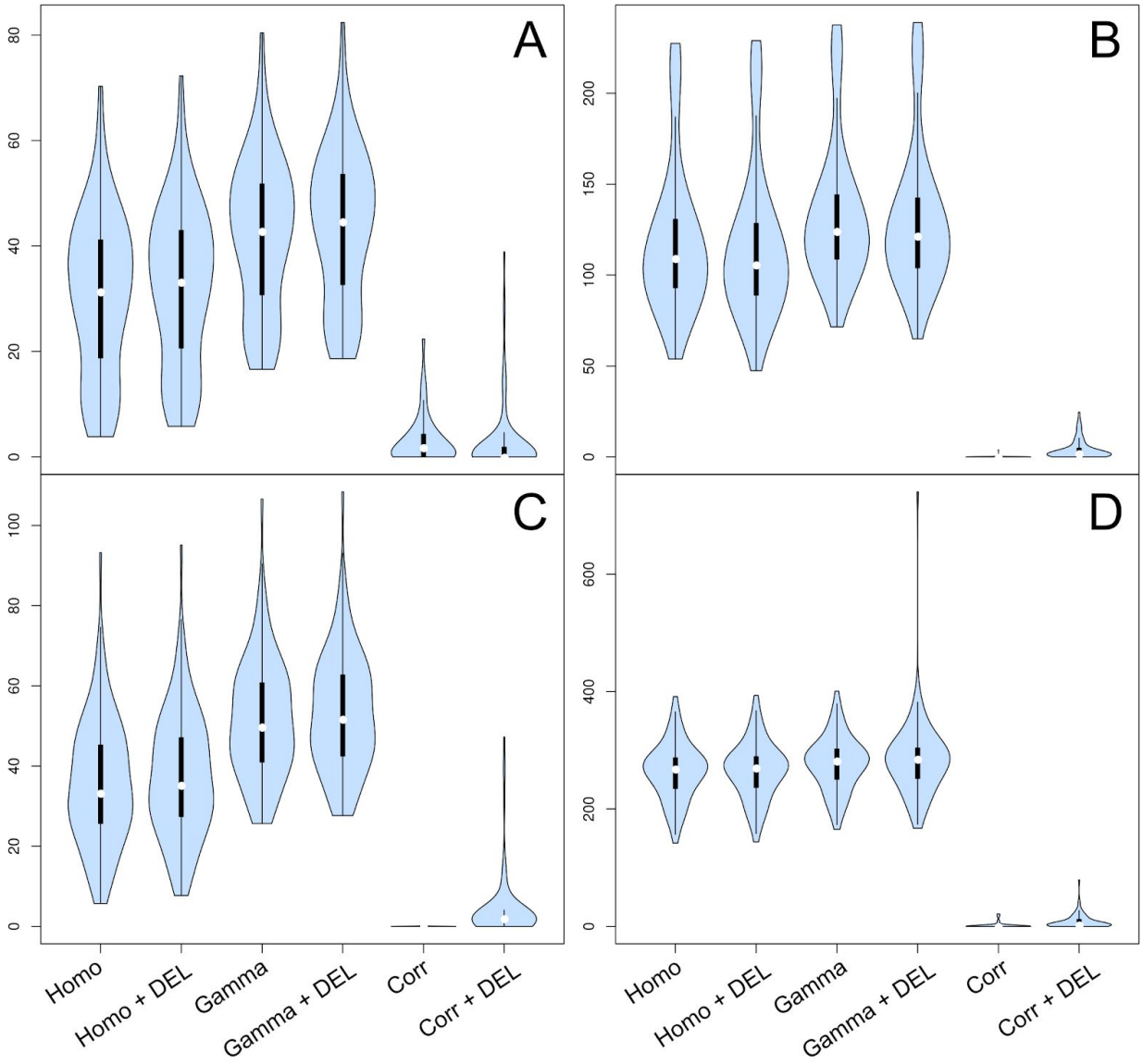

Figure S13: Distribution of delta Akaike Information Criterion (ΔAIC) for each model of sequence evolution applied to a single song bout for each species of *Gryllus* (see Figure 3 and Table 2 of the main text). A) Original sequences and distance-based alignments. B) Rescaled sequences and distance-based alignments. C) Original sequences and model-based alignments. D) Rescaled sequences and model-based alignments. See Table 2 for a description of the models.

2.7) Testing the robustness of results

In order to test the robustness of our results we simulated datasets with the same dimension of the *Gryllus* male calling song sequence data. We generated 100 ultrametric phylogenies with 12 extant species and tree height of 1 time unit. We used independent Markov models with the same parameter values we chose for the extended simulation study to generate phenotypic sequences with 728 positions for each phylogeny---same as the alignment of the observed data. We did not replicate the simulations using the original sequence data (Robillard et al. 2006) because it is not straightforward to simulate sequence traits of different length between species. We set the number of states to be fixed and equal to 2 across sequence positions---the state "A" represents sound elements and "G" silence elements.

We performed sequence alignment with the distance-based approach as implemented in MAFFT and using the same score matrix shown in Table S1. We did not align the sequence using BAli-Phy because of the significant computing resources necessary for the analyses coupled with the lack of a built-in signal to stop the MCMC sampler at a pre-selected number of MCMC generations. These characteristics of BAli-Phy generate non-trivial problems to the automatization of the pipeline across a large sample of data sets. For instance, one could use a parallel running program in order to automatically verify a stopping condition given the data on BAli-Phy output files and halt the BAli-Phy run by pairing the correct running process. However, such a program is beyond the objectives of the present study. Finally, we conducted model fit and model comparison using the Akaike Information Criterion.

The distribution of alignment lengths for the simulated data is similar to the empirical results (Figure S14), meaning that we can safely compare results of the simulation with the empirical data. The proportion of times that the correlated model was selected when it was not the generating model is acceptable (Figure S15 and Table S2). In contrast, the power of the test is relatively low, with only about 20% of the simulation replicates returning the correct model when the correlated model was the generating model. The low Type I error recovered by these simulations means that the chance the *Gryllus* sequences returned spurious evidence for correlated rates of evolution among sequence positions is acceptably low. Thus, we are confident about the results despite the low power due to the reduced size of the data set. Please refer to our discussion on the main text.

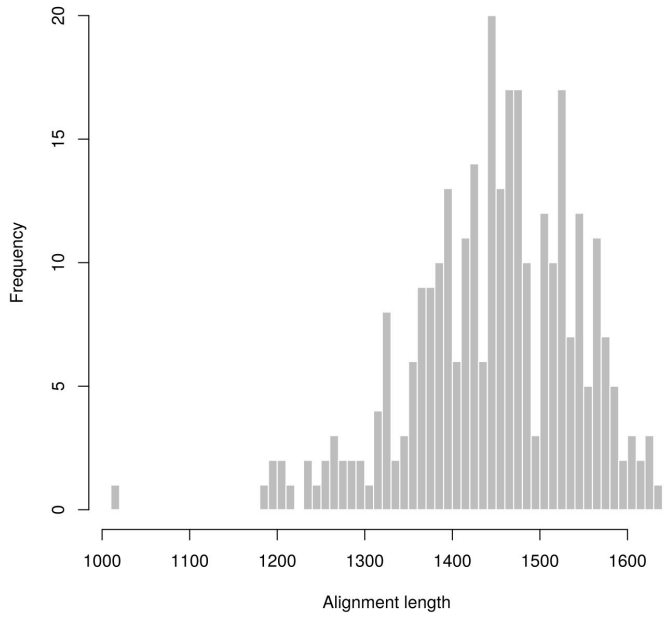

Figure S14: Distribution of alignment length for sequences simulated with the same dimension of the empirical complete song sequence data (12 species and 728 positions).

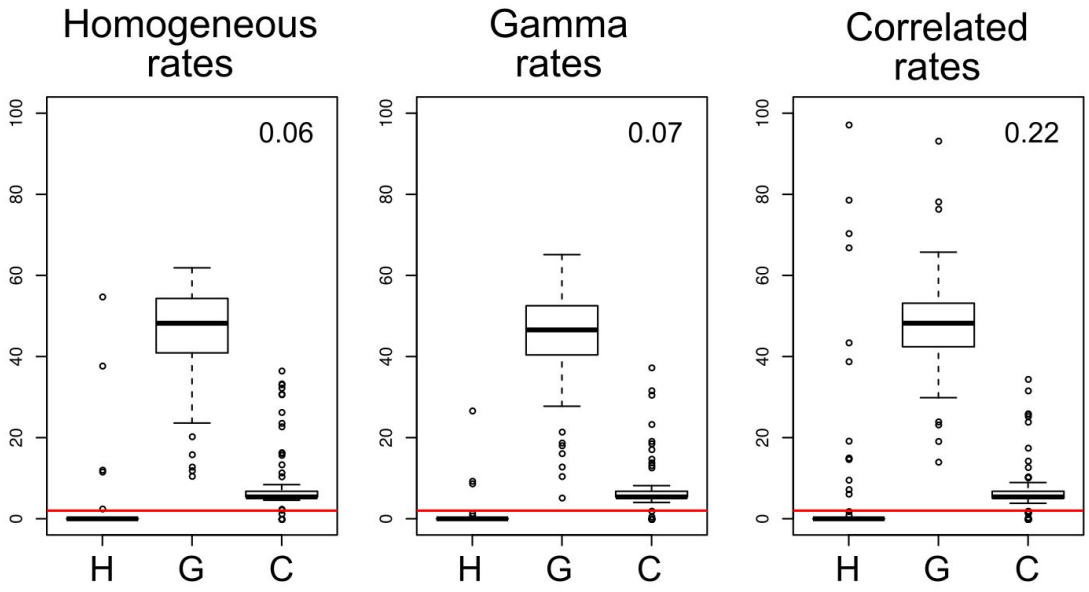

Figure S15: Distribution of pairwise ΔAIC across 100 simulation replicates for each of the three generating models with data sets showing the same number of species and sequence positions of the empirical data. H: homogeneous rates model; G: independent (Gamma) rates model; C: correlated rates model. Horizontal red lines mark the ΔAIC = 2 threshold. Numbers at the top right corner of each plot shows the proportion of times the correlated model was below the ΔAIC = 2 threshold, independent of the generating model (see also Table S3).

Table S2: Proportion of support for each of the models across 100 simulation replicates for each of the three generating models with data sets showing the same number of species and sequence positions of the empirical data. Models were considered supported if ΔAIC < 2 units (also shown as the red horizontal line on Figure S15).

Fitted models	**Generating (true) models**		
	Homogeneous	Gamma	Correlated
Homogeneous	0.95	0.97	0.86
Gamma	0.00	0.00	0.00
Correlated	0.06	0.07	0.22

2.8) Model-averaged transition rates estimated across sequence positions

Figures below show the distribution of transition rates estimated along the sequence positions of the alignments. These are complementary to Figure 2 and 3 of the main text. Figures 2 and 3 show only one single alignment as a representative of the results whereas here we summarize the variance of model-averaged transition rates estimated along sequence positions. The results shown below include phylogenetic uncertainty (Figure S16), alignment uncertainty (Figure S17), as well as variation of song bout (Figure S18 and S19).

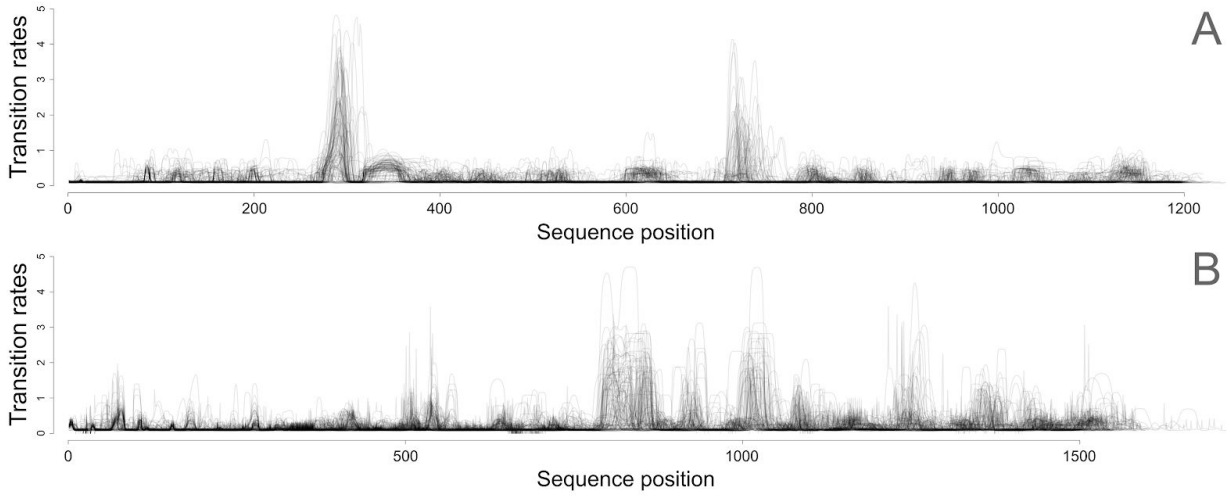

Figure S16: Model averaged rates of evolution estimated for the complete song sequences of *Gryllus* species under the correlated model using MAFFT conditioned on a random sample of 100 trees. A: Untransformed sequences. B: Rescaled sequences.

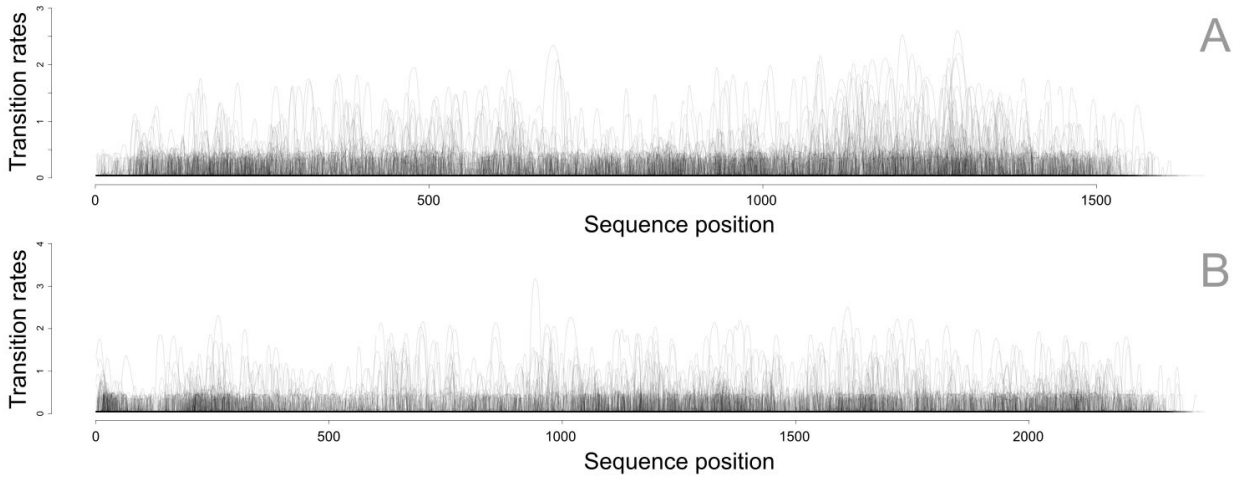

Figure S17: Model averaged rates of evolution estimated for the complete song sequences of *Gryllus* species using BAli-Phy conditioned on the MCC tree. Plots show random samples of 100 alignments from the posterior distribution. A: Untransformed sequences B: Rescaled sequences.

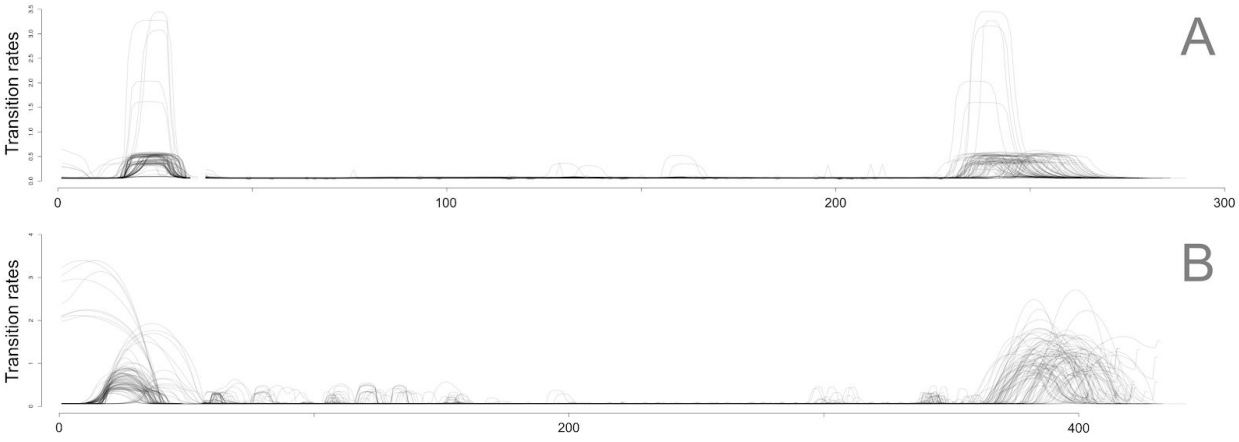

Figure S18: Model averaged rates of evolution estimated from a sample of 100 song bout data sets representing within species variation of songs. Alignments were estimated using MAFFT conditioned on the MCC tree. A) Original bout sequences (DO2). B) Time-rescaled bout sequences.

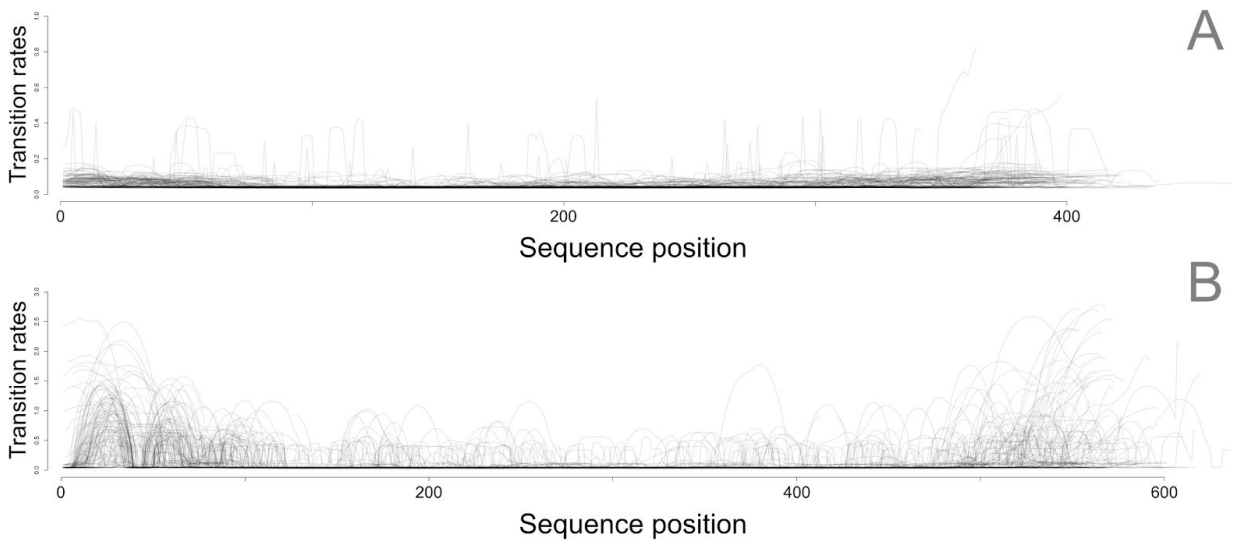

Figure S19: Rates of evolution estimated for the song sequence with a single song bout per species. Different from Figure S18, here only a single song bout data was used and the distribution of estimates reflects alignment uncertainty described by the posterior distribution from BAli-Phy. A) Original bout sequences (DO2). B) Time-rescaled bout sequences.

2.9) Comparing model-averaged rates between chirp and interchirp silence regions

In order to compare rates of transition estimated for chirp regions and interchirp silence regions we selected and grouped positions of the alignment between the groups. We selected a given alignment position as belonging to a chirp region (i.e., the column of the alignment) if more than 50% of the lineages (i.e., rows of the alignment) showed a sound character state. Likewise, we selected silence regions as positions with more than 50% silence character states. Then we recorded the distribution of model-averaged log transition rates for the positions of each group and computed median rate for each group. Here we used the median because the distribution of rates does not follow a Gaussian distribution. We repeated this operation across 100 randomly sampled alignments from the posterior distribution of the BAli-Phy analyses (i.e., incorporating alignment uncertainty) using both the time rescaled results for the whole sequence and the song bouts. We compared the median for the groups across samples using a paired Wilcoxon signed-rank test.

References

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C-H., Xie, D., Suchard, MA., Rambaut, A., & Drummond, A. J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Computational Biology, 10(4), e1003537. doi:10.1371/journal.pcbi.1003537

Darriba, D., G. L. Taboada, R. Doallo, and D. Posada. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Meth 9:772–772.

FitzJohn, R. G. 2012. Diversitree: comparative phylogenetic analyses of diversification in R. Methods in Ecology and Evolution 3:1084–1092

Gelman, A., and D. B. Rubin. 1992. Inference from Iterative Simulation Using Multiple Sequences. Statist. Sci. 7:457–472..

Henikoff, S., and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U.S.A. 89:10915–10919.

Huang, Y., G. Ortí, M. Sutherlin, A. Duhachek, and A. Zera. 2000. Phylogenetic Relationships of North American Field Crickets Inferred from Mitochondrial DNA Data. Molecular Phylogenetics and Evolution 17:48–57.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of Protein Molecules. Pp. 21–132 in Mammalian Protein Metabolism. Elsevier.

Katoh, K., and D. M. Standley. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 30:772–780.

Martyn Plummer, Nicky Best, Kate Cowles and Karen Vines (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC, R News, vol 6, 7-11

Otte, D. 1992. Evolution of Cricket Songs. Journal of Orthoptera Research 25.

Rambaut A, Drummond AJ, Xie D, Baele G and Suchard MA (2018) Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. Systematic Biology. syy032. doi:10.1093/sysbio/syy032

Redelings, B. D., and M. A. Suchard. 2005. Joint Bayesian Estimation of Alignment and Phylogeny. Systematic Biology 54:401–418.

Robillard, T., F. Legendre, L. Desutter-Grandcolas, and P. Grandcolas. 2006. Phylogenetic analysis and alignment of behavioral sequences by direct optimization. Cladistics 22:602–633.

Warren D.L., A.J. Geneva, and R. Lanfear. 2017. RWTY (R We There Yet): An R package for examining convergence of Bayesian phylogenetic analyses. Molecular Biology and Evolution 34:1016-1020. doi: 10.1093/molbev/msw279