RWTY (R We There Yet): An R Package for Examining Convergence of Bayesian Phylogenetic Analyses

Dan L. Warren,*[1] Anthony J. Geneva,[2] and Robert Lanfear[1,3]

[1]Department of Biological Sciences, Macquarie University, Sydney, Australia

[2]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA

[3]Division of Evolution, Ecology, and Genetics, Australian National University, Canberra, Australia

***Corresponding author:** E-mail: dan.l.warren@gmail.com.

Associate Editor: Michael Rosenberg

Abstract

Bayesian inference using Markov chain Monte Carlo (MCMC) has become one of the primary methods used to infer phylogenies from sequence data. Assessing convergence is a crucial component of these analyses, as it establishes the reliability of the posterior distribution estimates of the tree topology and model parameters sampled from the MCMC. Numerous tests and visualizations have been developed for this purpose, but many of the most popular methods are implemented in ways that make them inconvenient to use for large data sets. RWTY is an R package that implements established and new methods for diagnosing phylogenetic MCMC convergence in a single convenient interface.

***Key words:* MCMC, phylogenetics, topology, MCMC, convergence, stationarity.**

Bayesian inference using Markov chain Monte Carlo (MCMC) in phylogenetics involves inferring posterior distributions (e.g. of phylogenetic trees and model parameters) given a set of prior beliefs and a molecular sequence alignment. This approach has become one of the primary methods used to infer phylogenies from molecular data (Ronquist et al. 2012; Bouckaert et al. 2014). However, practical applications of MCMC to phylogenetic problems are complicated by a problem inherent to MCMC methods; it is frequently difficult to determine whether the chain has undergone enough iterations and whether enough samples have been taken to accurately infer the posterior distributions of clades and model parameters.

MCMC methods allow researchers to make inferences about the parameters of interest, such as the phylogenetic tree topology, while integrating out the uncertainty in other parameters, such as the model of molecular evolution (Gilks et al. 1996). It is not necessary to explore the entire space of possible solutions to do this; rather an MCMC chain is said to have "converged" when further exploration of the solution space does not change the inferred posterior probability distributions beyond some user-specified tolerance. Failure to appropriately diagnose non-convergence can lead to premature termination of chains, resulting in inappropriate estimates of the tree topology, clade support values, and model parameters.

There are attributes of the phylogeny problem that make achieving and assessing convergence difficult. Phylogenetic problems often involve estimating many interacting model parameters (such as rates of evolution and dates of divergence), as well as the topology of the phylogenetic tree (which may itself interact with inferred model parameters). Interactions among continuous parameters can make exploring the space of possible solutions difficult, because efficient exploration requires coordinated changes among more than one parameter. On top of this, exploring the space of all possible phylogenetic tree topologies ("tree space") is also difficult; the number of possible tree topologies is astronomical even for small numbers of taxa, and adjacent solutions can differ considerably in their posterior probability. As such, tree space can contain local optima (Whidden and Matsen 2015) in which the MCMC can become stuck, and so fail to converge. Interactions among parameters mean that the failure of a single parameter to converge can lead to poor inferences of other parameters. Thus, assessing MCMC convergence requires the analysis of all parameters, including the tree topology itself.

Initially, MCMC convergence in phylogenetics was primarily diagnosed using plots of log likelihood as a function of chain length (fig. 1, panel *C*). Although a converged chain will have a relatively flat likelihood trace, this is not a sufficient condition for diagnosing chain convergence; a chain that is stuck on a single local optimum will also produce a relatively flat likelihood trace while potentially being far from convergence. Similarly, a chain that is exploring multiple local optima with similar likelihoods may produce a flat likelihood trace without producing accurate posterior probability distributions. More recently, these methods have been extended, and it is now standard practice to assess convergence by examining the traces and posterior distributions of all continuous parameters in the analysis (Höhna and Drummond 2012; Rambaut et al. 2014). However, these approaches still fail to address a key question: whether or not the MCMC has adequately sampled the space of potential topologies.

In order to deal with the shortcomings of convergence diagnostics based on likelihoods and model parameters, AWTY (Are We There Yet, Nylander et al. 2008) provided novel convergence diagnostics based on directly examining

doi:10.1093/molbev/msw279 Advance Access publication January 19, 2017

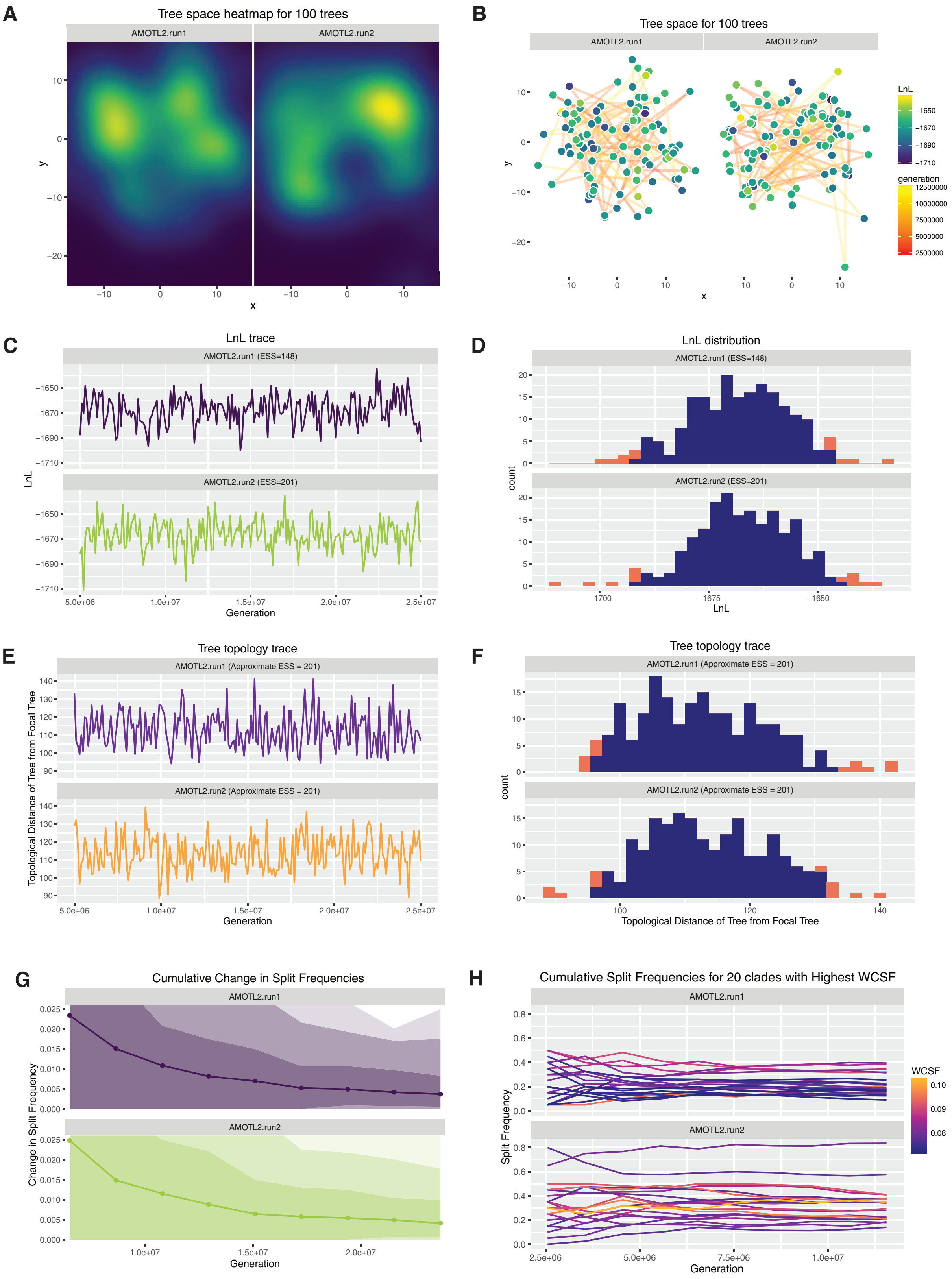

Fig. 1. RWTY plots examining the behavior of individual chains. Data are two chains from Williams et al. (2013). RWTY allows users to visualize the amount of time chains have spent in different areas of tree space both as a heatmap (*A*) and a scatter plot (*B*). RWTY also allows users to visualize likelihoods and model parameters (*C* and *D*) and tree topology (*E* and *F*) as a function of chain length. In panel *G*, the cumulative change in split frequencies is plotted as a function of chain length, where the solid line gives the mean standard deviation of split frequencies as a function of chain length and increasingly lighter ribbons give the limits of the 75%, 95%, and 100% quantiles. RWTY also plots the cumulative posterior probability estimates for clades as a function of chain length, and highlights splits that are likely to be problematic using a metric (WCSF) that weights changes in split frequencies by their position in the chain (H).

the posterior probabilities of clades as a function of chain length. These diagnostics help users detect when multiple topological optima are being explored, and can help estimate the number of trees necessary to achieve accurate posterior estimates (fig. 1, panel *H*). Comparison of such plots from multiple replicate chains can further assist users in diagnosing problems with phylogenetic MCMC analyses because well-behaved replicate chains will infer similar posterior distributions. It must be noted that none of these diagnostic plots is sufficient to positively diagnose convergence, but at minimum they represent a much stricter set of necessary conditions for accepting the output of an MCMC when compared to simply examining likelihood or parameter plots alone.

The AWTY software (Nylander et al. 2008) has been widely used since its release, and is now part of the standard toolbox of investigators using MCMC methods in phylogenetics. However, the package is only available through an online interface, it runs slowly on large tree files, it does not read tree files from the latest phylogenetic MCMC software, and its code is not open source and so cannot be extended or further developed by the community.

RWTY is a new package that leverages the functionality of R packages for phylogenetics (Paradis et al. 2004; Schliep 2011), statistical analysis (Plummer et al. 2006; Wickham 2007; Schloerke et al. 2011; R Core Team 2016), and visualization (Wickham 2009; Schloerke et al. 2011; de Vries and Ripley 2013; Garnier 2016) to provide a suite of functions for visualizing and analyzing the performance of MCMC chains. RWTY provides a single environment in which to analyze the convergence of all parameters in a phylogenetic MCMC analysis, including continuous parameters and those associated with the tree topology. RWTY accepts input from popular phylogenetic MCMC packages, currently including MrBayes (Ronquist et al. 2012), BEAST (Bouckaert et al. 2014), and RevBayes (Höhna et al. 2016). In addition, trees may be manually loaded from any format that can be coerced into an ape multiphylo object, and parameter data from any format that can be converted to an R data frame. RWTY provides access to many existing and new methods for assessing convergence of phylogenetic MCMC analyses. For example, it produces plots of the posterior probability of sampled clades similar to those produced by AWTY (Nylander et al. 2008) (fig. 1, panel *H*), it allows visualization of MCMC exploration of tree space in a manner similar to TreeSetViz (Amenta and Klingner 2002) (fig. 1, panels *A* and *B*), and it allows users to examine traces, posterior probability distributions, and effective sample sizes (ESS) of model parameters similar to Tracer (Rambaut et al. 2014) (fig. 1, panels *C*–*F*).

RWTY also implements several new methods that focus specifically on assessing the adequacy with which the MCMC has sampled the phylogenetic tree topology space. These include new visualizations of the trace and distribution of tree topologies sampled by the MCMC (Lanfear et al. 2016), visualizations of the similarity of tree topologies sampled by different chains (fig. 2), visualizations of changes in split frequencies within chains and differences in split frequencies between chains as the MCMC progressed, methods to

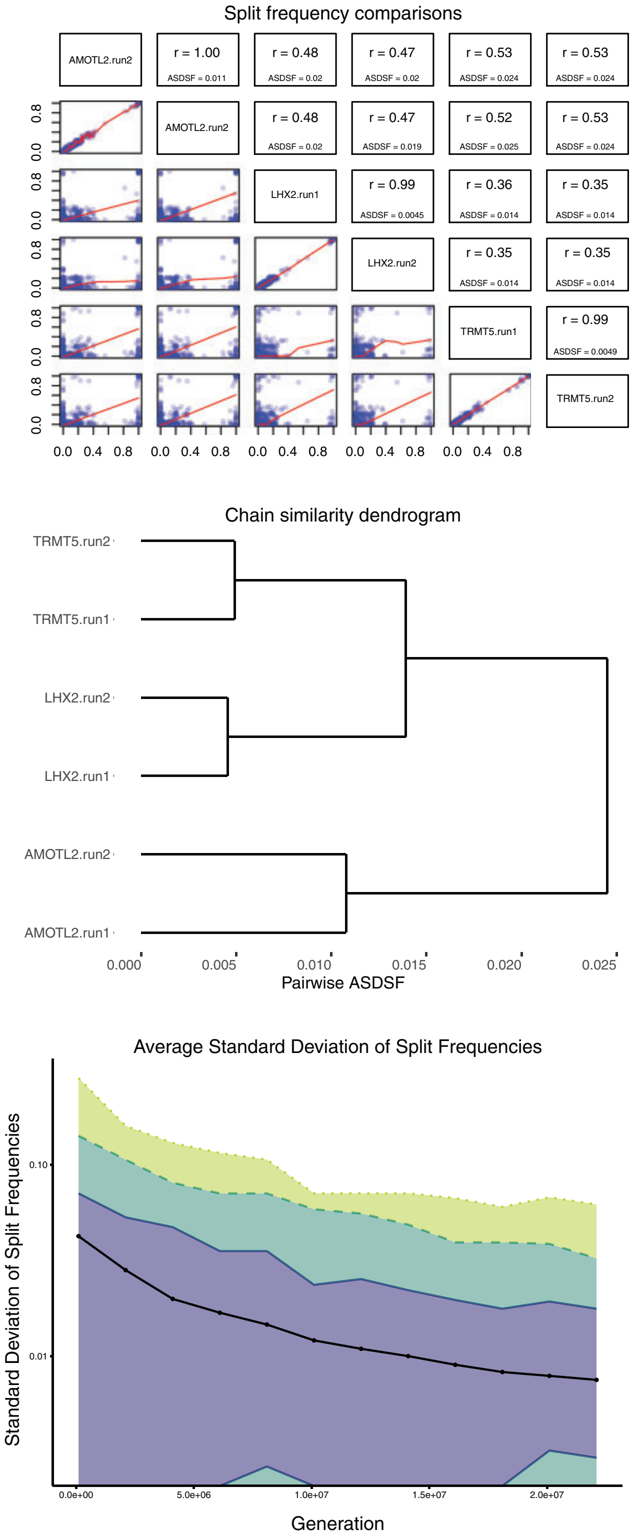

Fig. 2. RWTY plots for between-chain comparisons. Data are six MCMC chains from an analysis of three loci, two runs per locus (Williams et al. 2013). In the top panel, posterior probability estimates from each pair of chains are shown in a scatter plot (below diagonal), and summary statistics are given above the diagonal. The center panel shows a chain comparison dendrogram, in which the branch length separating each pair of chains represents the average standard deviation of split frequencies between those chains. In the lower panel, the mean and 75%, 95% and 100% quantiles of the standard deviation of split frequencies across all chains is plotted against chain length.

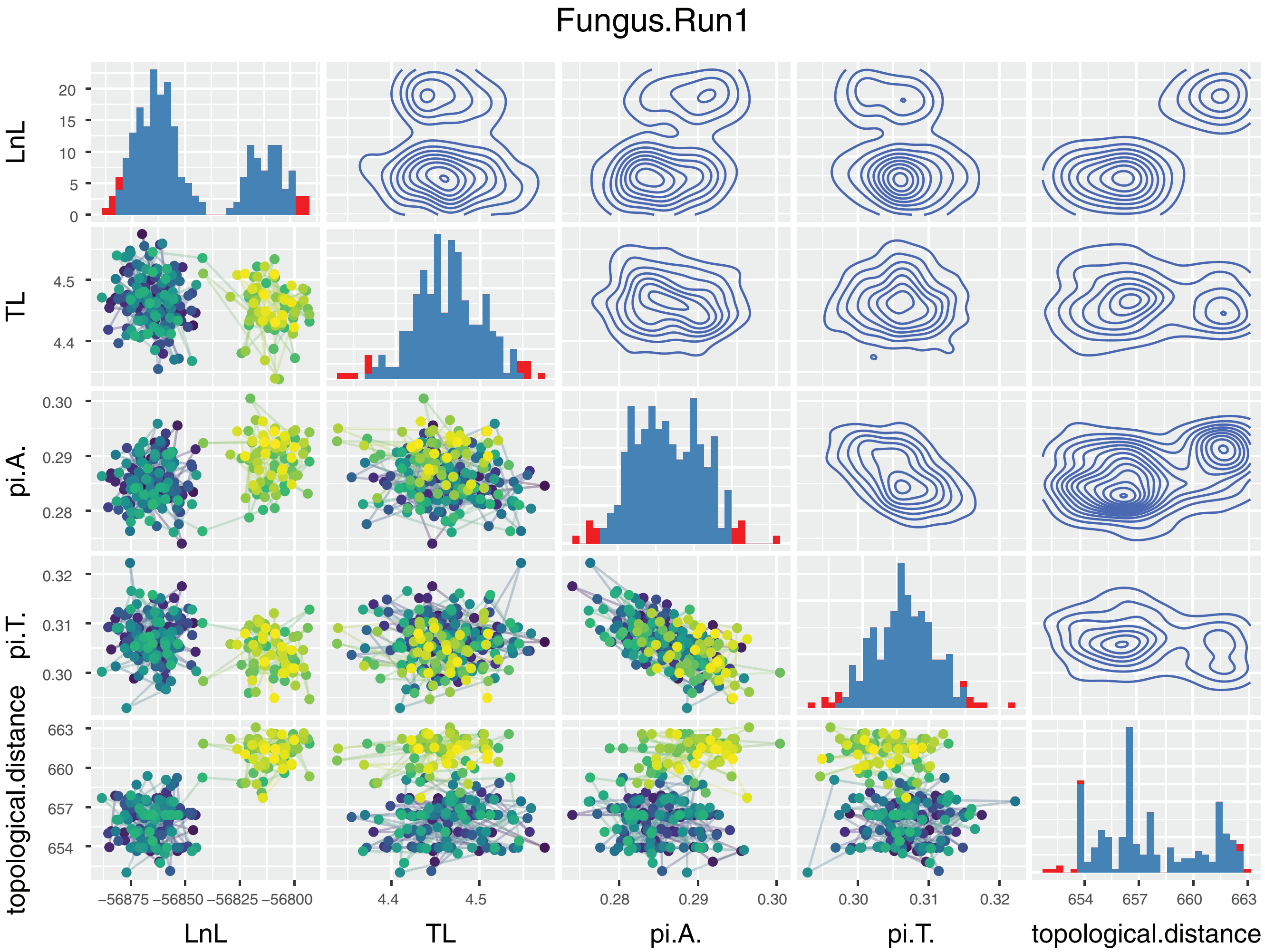

Fig. 3. Pairwise examination of topology, model parameters, and likelihoods for a single chain from an analysis of a fungus dataset (Hibbett et al. 1997). Histograms along the diagonal plot the posterior probability distribution for each parameter, with red indicating values outside of the 95% confidence interval. Relationships between parameters are displayed as scatter plots below the diagonal and contour plots above the diagonal. Points in the scatter plots are colored according to generation from the MCMC chain, with lighter colors representing points later in the chain.

calculate the ESS of the tree topologies sampled (Lanfear et al. 2016), and visualizations of the autocorrelation of tree topologies sampled from each chain (Lanfear et al. 2016). All functionality can be accessed directly using single-purpose functions, but users will typically interact with RWTY using a single omnibus function, analyze.rwty. The analyze.rwty function automatically determines which plots are possible with the provided data, and produces an R object containing all plots. Examples of some plot types output by RWTY are presented in figs. 1–3. The software and a comprehensive vignette (supplementary data S1 and S2, Supplementary Material online) are available at https://github.com/danlwarren/RWTY, or from the Comprehensive R Archive Network (https://cran.r-project.org/). R users with an internet connection can install the package using the command “install.packages (‘rwty’)” or “library(devtools); install_github (‘danlwarren/RWTY’)”.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

This work was funded in part by Australian Research Council fellowship awards to D.L.W. and R.L.

References

Amenta N, Klingner J. 2002. Case study: visualizing sets of evolutionary trees. In: IEEE symposium on information visualization, 2002. INFOVIS 2002. Washington (DC): IEEE. p. 71–74.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol*. 10:3537.

de Vries A, Ripley B. 2013. Ggdendro: tools for extracting dendrogram and tree diagram plot data for use with ggplot. R package version 0.1-12. Available from: http://cran/.R-project.org/package=ggdendro.

Garnier S. 2016. viridis: default color maps from ′matplotlib′. R package version 0.3.4.

Gilks WR, Richardson S, Spiegelhalter DJ. 1996. Introducing Markov chain Monte Carlo. *Markov Chain Monte Carlo Practice* 1:19.

Hibbett DS, Pine EM, Langer E, Langer G, Donoghue MJ. 1997. Evolution of gilled mushrooms and puffballs inferred from ribosomal DNA sequences. *Proc Natl Acad Sci*. 94:12002–12006.

Höhna S, Drummond AJ. 2012. Guided tree topology proposals for Bayesian phylogenetic inference. *Syst Biol*. 61:1–11.

Höhna S, Landis M, Heath T, Boussau B, Lartillot N, Moore B, Huelsenbeck J, Ronquist F. 2016. RevBayes: a flexible framework for Bayesian inference of phylogeny. *Syst Biol*. 64:726–736.

Lanfear R, Hua X, Warren DL. 2016. Estimating the effective sample size of tree topologies from Bayesian phylogenetic analyses. *Genome Biol Evol*. 8:2319–2332.

Nylander JA, Wilgenbusch JC, Warren DL, Swofford DL. 2008. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics* 24:581–583.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290.

Plummer M, Best N, Cowles K, Vines K. 2006. CODA: convergence diagnosis and output analysis for MCMC. *R News*. 6:7–11.

R Core Team. 2016. R: a language and environment for statistical computing. Vienna, Austria. Available from: http://cran/.R-project.org/.

Rambaut A, Suchard M, Xie D, Drummond A. 2014. Tracer v1. 6. Available from: http://tree.bio.ed.ac.uk/software/tracer/.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*. 61:539–542.

Schliep KP. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27:592–593.

Schloerke B, Crowley J, Cook D, Hofmann H, Wickham H, Briatte F, Marbach M, Thoen E. 2011. Ggally: extension to ggplot2.

Whidden C, Matsen FA. 2015. Quantifying MCMC exploration of phylogenetic tree space. *Syst Biol*. 64:472–491.

Wickham H. 2009. ggplot2: elegant graphics for data analysis. New York: Springer Science & Business Media.

Wickham H. 2007. Reshaping data with the reshape package. *J Stat Softw*. 21:1–20.

Williams JS, Niedzwiecki JH, Weisrock DW. 2013. Species tree reconstruction of a poorly resolved clade of salamanders (Ambystomatidae) using multiple nuclear loci. *Mol Phylogenet Evol*. 68:671–682.