Software for Systematics and Evolution

Syst. Biol. 67(5):901–904, 2018

DOI:10.1093/sysbio/syy032
Advance Access publication April 27, 2018

Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7

ANDREW RAMBAUT [1,*], ALEXEI J. DRUMMOND [2,3], DONG XIE [2,3], GUY BAELE [4], AND MARC A. SUCHARD [5,6]

[1]*Institute of Evolutionary Biology, University of Edinburgh, Ashworth Laboratories, King's Buildings, Edinburgh, EH9 3FL, UK;* [2]*Department of Computer Science, University of Auckland, 303/38 Princes St, Auckland, 1010, NZ;* [3]*Centre for Computational Evolution, University of Auckland, 303/38 Princes St, Auckland, 1010, NZ;* [4]*Department of Microbiology and Immunology, Rega Institute, KU Leuven - University of Leuven, Herestraat 49, 3000 Leuven, Belgium;* [5]*Department of Human Genetics, University of California, Los Angeles, 695 Charles E. Young Dr., Los Angeles, CA 90095, USA; and* [6]*Department of Biostatistics, University of California, Los Angeles, 650 Charles E. Young Dr., Los Angeles, CA 90095, USA*
[*]*Correspondence to be sent to: Institute of Evolutionary Biology, University of Edinburgh, Ashworth Laboratories, King's Buildings, Edinburgh, EH9 3FL, UK;*
Email: a.rambaut@ed.ac.uk.

Received 13 February 2018; reviews returned 19 April 2018; accepted 20 April 2018
Associate Editor: Edward Susko

Abstract.— Bayesian inference of phylogeny using Markov chain Monte Carlo (MCMC) plays a central role in understanding evolutionary history from molecular sequence data. Visualizing and analyzing the MCMC-generated samples from the posterior distribution is a key step in any non-trivial Bayesian inference. We present the software package *Tracer* (version 1.7) for visualizing and analyzing the MCMC trace files generated through Bayesian phylogenetic inference. *Tracer* provides kernel density estimation, multivariate visualization, demographic trajectory reconstruction, conditional posterior distribution summary, and more. *Tracer* is open-source and available at `http://beast.community/tracer`. [Bayesian inference; Markov chain Monte Carlo; phylogenetics; visualization]

Bayesian inference of phylogeny using Markov chain Monte Carlo (MCMC) (Rannala and Yang 1996; Mau et al. 1999; Drummond et al. 2002) flourishes as a popular approach to uncover the evolutionary relationships among taxa, such as genes, genomes, individuals, or species. MCMC approaches generate samples of model parameter values—including the phylogenetic tree—drawn from their posterior distribution given molecular sequence data and a selection of evolutionary models. Visualizing, tabulating, and marginalizing these samples are critical for approximating the posterior quantities of interest that one reports as the outcome of a Bayesian phylogenetic analysis. To facilitate this task, we have developed the *Tracer* (version 1.7) software package to process MCMC trace files containing parameter samples and to interactively explore the high-dimensional posterior distribution. *Tracer* works automatically with sample output from BEAST (Drummond et al. 2012), BEAST2 (Bouckaert et al. 2014), LAMARC (Kuhner 2006), Migrate (Beerli 2006), MrBayes (Ronquist et al. 2012), RevBayes (Höhna et al. 2016), and possibly other MCMC programs from other domains.

DESIGN AND IMPLEMENTATION

Tracer examines the posterior samples from all the available parameters—treating continuous, integer and categorical parameters appropriately—from a trace and presents statistical summaries and visualizations. Further, *Tracer* can analyze a single trace or combine samples from multiple files. Immediately apparent in the default *Tracer* view, the effective sample size (ESS) is one such statistic that allows users to assess the number of effectively independent draws from the posterior distribution the trace represents (Figure 1a). Color coding assists the user in determining potential MCMC mixing problems, with arbitrary cut-off values at 100 and 200.

Selecting multiple parameters from the "Traces" panel on the left generates a side-by-side comparison or an overlay of the selected parameters' visualizations (Figure 1 b–e). Multiple trace files can be selected in a similar fashion to compare posterior samples between different replicates of an analysis. If multiple trace files contain the same collection of parameters, then a "Combined" trace appears automatically. *Tracer* generates four display panels for the selected parameters:

- Estimates: Reports common summary statistics such as the sample mean, standard deviation, highest posterior density interval, and ESS. Also presents a histogram of sample values for a single selected parameter (Figure 1a) or side-by-side boxplots for multiple continuous parameters (Figure 1b).
- Marginal density: Draws density plots for the selected parameter(s), including kernel density estimates (Figure 1c), histograms, and violin plots (Figure 1d) for continuous parameters and frequency plots for categorical or integer parameters.
- Joint-marginal: Visualization in this panel appears after selecting two or more parameters, and the

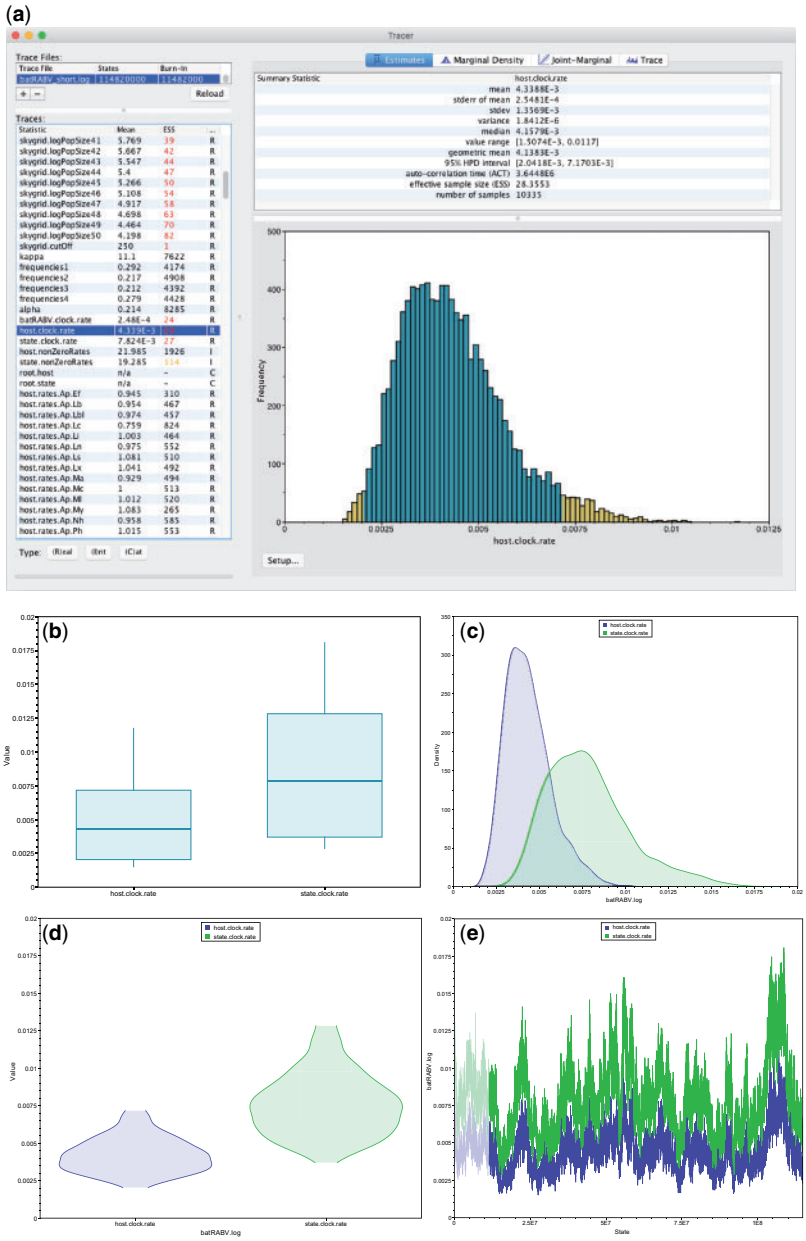

FIGURE 1. Overview of *Tracer* functionality and individual parameter visualizations: a) Main *Tracer* panel upon loading a single trace file; b) boxplot representation of two continuous parameters; corresponding c) kernel density estimates; d) violin plots; e) the actual traces connecting the parameter values visited by the Markov chain.

plot form depends on the parameter types. We show several examples in the next section of the article.

- Trace: Constructs line plots connecting the sequential samples of one or more selected parameters against state or generation number (Figure 1e). Users typically use this plot to assess mixing, select a suitable burn-in and identify trends that suggest convergence issues.

Tracer offers a solution of visualizing conditional posterior distributions as well. Selecting one continuous and one integer or categorical parameter generates side-by-side violin or boxplots under the Joint-Marginal panel. These plots present the continuous parameter distribution conditioned on the unique integer or categorical values. A typical use case involves Bayesian stochastic search variable selection (BSSVS), a form of model averaging, in which parameters influence the likelihood function only when a specific model is selected by a random indicator function. Under BSSVS, a posterior estimate of the parameter should only sample values from states where the indicator equals one. Discrete phylogeographic analyses frequently employ BSSVS due to the potentially large amount of transition rates that need to be estimated (Lemey et al. 2009), but

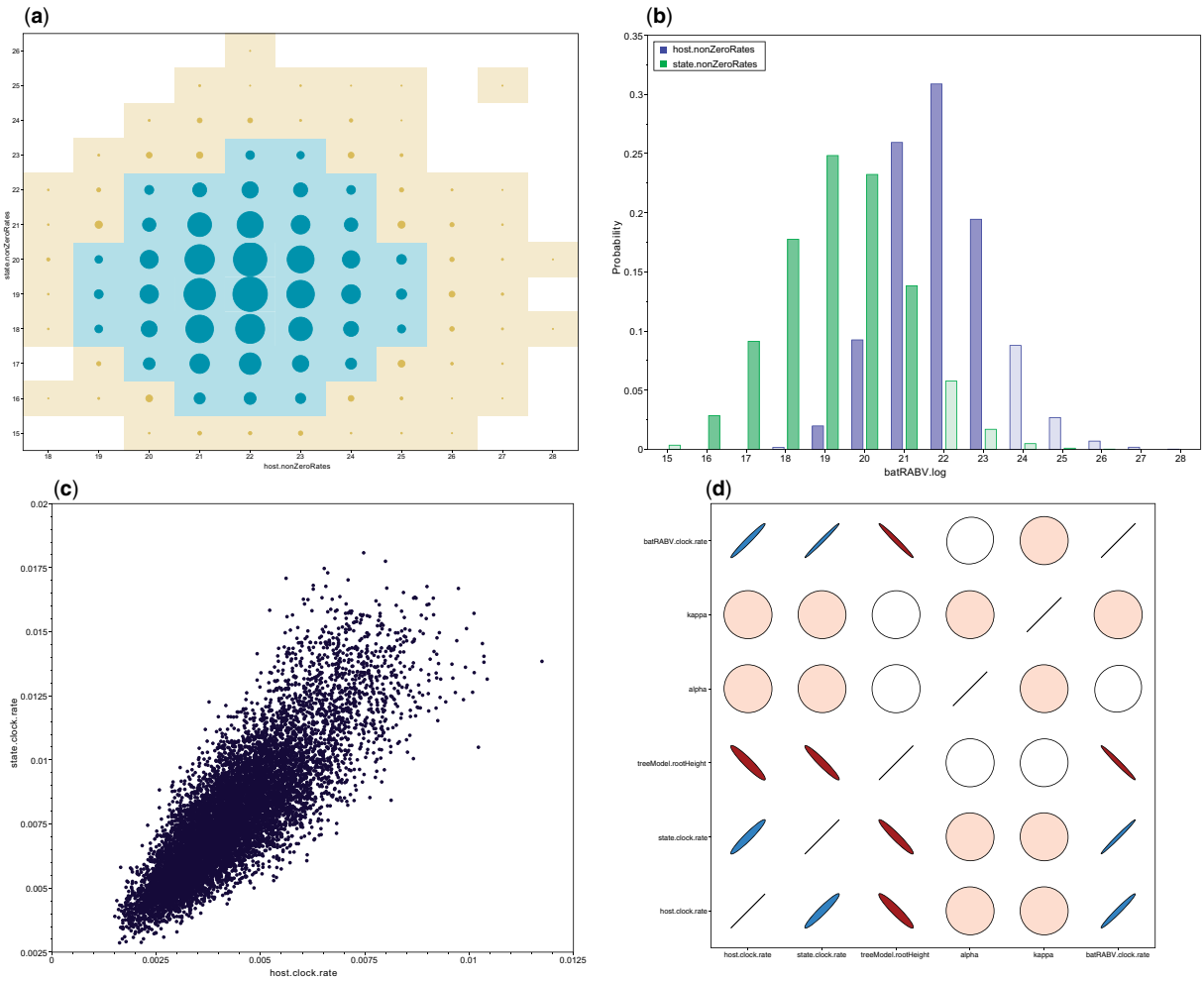

FIGURE 2. Multi-parameter visualizations of: a) the joint probability distribution of two integer variables through a bubble chart; b) the marginal density of multiple integer or categorical variables through frequency plots; c) two continuous variables through a classic scatter or correlation plot; d) multiple (> 2) continuous variables using large correlation matrices.

this is also relevant when employing model averaging approaches, e.g., over relaxed molecular clocks (Li and Drummond 2012).

Finally, *Tracer* provides demographic reconstruction resulting in a graphical plot, often applied to reconstruct epidemic dynamics. Available models include, e.g., constant size, exponential and logistic growth (Drummond et al. 2002), and the non-parametric Bayesian skyline (Drummond et al. 2005; Heled and Drummond 2008), skyride (Minin et al. 2008), and skygrid (Gill et al. 2013).

EXAMPLE

Cross-Species Dynamics of North American Bat Rabies

We use *Tracer* to infer the spatial dispersal and cross-species dynamics of rabies virus (RABV) in North American bats. The data set comprises 372 *nucleoprotein* gene sequences from 17 bat species, sampled between 1997 and 2006 across 14 states in the United States (Streicker et al. 2010; Faria et al. 2013). We estimate RABV ancestral locations and host-jumping history using a Bayesian discrete phylogeographic approach with BSSVS, while simultaneously estimating effective population sizes over time through a Bayesian skygrid coalescent model (Gill et al. 2013).

Phylogeographic BSSVS inference includes parameters of both integer (number of non-zero transition rates) and categorical (host or location-state) trace types. In *Tracer*, a bubble chart visualizes the joint probability distribution between two integer or categorical traces (see Figure 2a). Circle area is proportional to the joint probability, with a colored tile background if this probability reaches a nominal threshold to enhance visibility. Marginal density plots can also display multiple integer parameters, each with unique colour scales (see Figure 2b). With approximately equal numbers of transition rates, both figures suggest similar host and location trait model complexity. *Tracer* also provides popular visualizations for continuous parameters, including scatter plots for two parameters (see Figure 2c), and extensions for correlations between ≥2 continuous parameters (Figure 2d; Murdoch and Chow 1996). Colour gradients indicate strength and

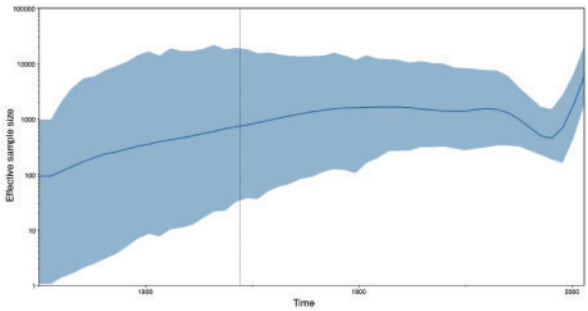

FIGURE 3. Estimating the effective population sizes over time using a Bayesian skygrid demographic reconstruction for rabies virus in North America.

direction of the correlation, from red (strong negative) to blue (strong positive). Ellipse shapes re-enforce the strength of correlation, with no correlation appearing as a circle and perfect (anti)correlation as a line.

Tracer reconstructs the demographic history of RABV by drawing the effective population sizes over time (Figure 3). RABV has successfully established itself in North American bat species, with its effective population size rising steadily throughout recent centuries. Following a rapid decline at the end of last century, we observe a recent sharp increase in size.

Other packages are available for the post-processing of MCMC samples. "coda" (Plummer et al. 2006) provides some of the functionality of *Tracer* within the R programming environment, while "AWTY" (Nylander et al. 2007) and "RWTY" (Warren et al. 2017) explore the convergence of the phylogenetic tree parameter itself across multiple MCMC runs. These alternative packages compute, e.g., Gelman–Rubin diagnostics (Gelman and Rubin 1992) that *Tracer* currently does not provide.

AVAILABILITY

Tracer is open-source under the GNU lesser general public license and available in both source code (https://github.com/beast-dev/tracer) and executable (http://beast.community/tracer) forms. This latter page also serves up self-contained, step-by-step tutorials covering basic to advanced usage of *Tracer* to summarize posteriors under a variety of phylogenetic models using BEAST and diagnose MCMC chain convergence. Popular tutorials employ *Tracer* to generate marginal parameter summaries and to infer population dynamics trajectories over time. *Tracer* requires Java version 1.6 or greater.

FUNDING

This work was supported in part by the Wellcome Trust through project 206298/Z/17/Z (Artic Network), the European Union Seventh Framework Programme under [grant agreement no. 725422-RESERVOIRDOCS]; the National Science Foundation through grant DMS [1264153]; the National Institutes of Health under grants [R01 AI107034 and U19 AI135995]; Marsden grant contract [UOA1611 to A.J.D.]; Interne Fondsen KU Leuven / Internal Funds KU Leuven to G.B.

REFERENCES

Beerli P. (2006). Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. Bioinformatics 22:341–345.

Bouckaert R., Heled J., Kühnert D., Vaughan T., Wu C.-H., Xie D., Suchard M.A., Rambaut A., Drummond A.J. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comp. Biol. 10:e1003537.

Drummond, A.J., Nicholls, G.K., Rodrigo, A.G., Solomon W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 161:1307–1320.

Drummond, A.J., Rambaut, A., Shapiro, B., Pybus, O.G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. 22:1185–1192.

Drummond A.J., Suchard M.A. Xie D., Rambaut A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969–1973.

Faria, N., Suchard M., Rambaut A., Streicker D., Lemey P. (2013). Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraint. Philos. Trans. R. Soc. Lon. B 368:20120196.

Gelman A., Rubin D.B. (1992). Inference from iterative simulation using multiple sequences. Stat. Sci. 7:457–472.

Gill, M.S., Lemey, P., Faria N.R., Rambaut A., Shapiro B., Suchard M.A. (2013). Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Mol. Biol. Evol. 30:713–724.

Heled J., Drummond A.J. (2008). Bayesian inference of population size history from multiple loci. BMC Evol. Biol. 8:289.

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

Kuhner M.K. (2006). LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics 22: 768–770.

Lemey P., Rambaut A., Drummond A., Suchard M. (2009). Bayesian phylogeography finds its roots. PLoS Comp. Biol. 5:e1000520.

Li W.L.S., Drummond A.J. (2012). Model averaging and Bayes factor calculation of relaxed molecular clocks in Bayesian phylogenetics. Mol. Biol. Evol. 29:751–761.

Mau B., Newton M.A., Larget B. (1999). Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics 55:1–12.

Minin V.N., Bloomquist E.W., Suchard M.A. (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol. Biol. Evol. 25:1459–1471.

Murdoch D., Chow E. (1996). A graphical display of large correlation matrices. Am. Stat. 50:178–180.

Nylander J.A., Wilgenbusch J.C., Warren D.L., Swofford D.L. (2007). AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 24:581–583.

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

Rannala B., Yang Z. (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J. Mol. Evol. 43:304–311.

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

Streicker D., Turmelle A., Vonhof M., Kuzmin I., McCracken G.F., Rupprecht C. (2010). Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. Science 329:676–679.

Warren D.L., Geneva A.J., Lanfear R. (2017). RWTY (R We There Yet): an R package for examining convergence of Bayesian phylogenetic analyses. Mol. Biol. Evol. 34:1016–1020.