Taming the BEAST—A Community Teaching Material Resource for BEAST 2

Joëlle Barido-Sottani[1,2,†], Veronika Boškova[1,2,†], Louis du Plessis[1,3,†], Denise Kühnert[1,2,4,†], Carsten Magnus[1,2,†], Venelin Mitov[1,2,†], Nicola F. Müller[1,2,†], Jūlija Pečerska[1,2,†], David A. Rasmussen[1,2,†], Chi Zhang[1,2,†], Alexei J. Drummond[5,‡], Tracy A. Heath[6,‡], Oliver G. Pybus[3,‡], Timothy G. Vaughan[5,‡], and Tanja Stadler[1,2,*,§]

[1]*Department of Biosystems Science and Engineering, ETH Zürich, Mattenstrasse 26, 4058 Basel, Switzerland;* [2]*Swiss Institute of Bioinformatics (SIB), Quartier Sorge - Batiment Genopode, 1015 Lausanne, Switzerland;* [3]*Department of Zoology, University of Oxford, Peter Medawar Building South Parks Road Oxford, OX1 3SY, UK;* [4]*Department of Environmental Sciences, ETH Zürich, Universitätsstrasse 16, 8092 Zürich, Switzerland;* [5]*Centre for Computational Evolution, University of Auckland, New Zealand; and* [6]*Department of Ecology, Evolution, and Organismal Biology, Iowa State University, 2200 Osborn Dr., Ames, IA 50011 USA*
*Correspondence to be sent to: Department of Biosystems Science and Engineering, ETH Zürich, Mattenstrasse 26, 4058 Basel, Switzerland; E-mail: tanja.stadler@bsse.ethz.ch.
†These authors contributed equally. ‡These authors contributed equally.
§Senior author.

Abstract.—Phylogenetics and phylodynamics are central topics in modern evolutionary biology. Phylogenetic methods reconstruct the evolutionary relationships among organisms, whereas phylodynamic approaches reveal the underlying diversification processes that lead to the observed relationships. These two fields have many practical applications in disciplines as diverse as epidemiology, developmental biology, palaeontology, ecology, and linguistics. The combination of increasingly large genetic data sets and increases in computing power is facilitating the development of more sophisticated phylogenetic and phylodynamic methods. Big data sets allow us to answer complex questions. However, since the required analyses are highly specific to the particular data set and question, a black-box method is not sufficient anymore. Instead, biologists are required to be actively involved with modeling decisions during data analysis. The modular design of the Bayesian phylogenetic software package BEAST 2 enables, and in fact enforces, this involvement. At the same time, the modular design enables computational biology groups to develop new methods at a rapid rate. A thorough understanding of the models and algorithms used by inference software is a critical prerequisite for successful hypothesis formulation and assessment. In particular, there is a need for more readily available resources aimed at helping interested scientists equip themselves with the skills to confidently use cutting-edge phylogenetic analysis software. These resources will also benefit researchers who do not have access to similar courses or training at their home institutions. Here, we introduce the "Taming the Beast" (https://taming-the-beast.github.io/) resource, which was developed as part of a workshop series bearing the same name, to facilitate the usage of the Bayesian phylogenetic software package BEAST 2. [Bayesian inference; MCMC; phylodynamics; phylogenetics.]

BEAST 2 in a Nutshell

BEAST 2 (Bouckaert et al., 2014) is an open source cross-platform software package for analysing genetic sequences in a Bayesian phylogenetic framework. It occupies the same niche, and thus incorporates many of the same models, as other popular Bayesian evolutionary analyses platforms, including BEAST (Drummond and Rambaut, 2007) (which we refer to here as BEAST 1 in order to distinguish it from BEAST 2), MrBayes (Huelsenbeck and Ronquist, 2001), and RevBayes (Höhna et al., 2016). Although BEAST 2 is a complete redesign of the BEAST 1 software package, it retains a similar user interface and many core model components, including relaxed molecular clock models (Drummond et al., 2006), Bayesian skyline models for nonparametric coalescent analyses (Drummond et al., 2005; Heled and Drummond, 2008), multispecies coalescent inference with *BEAST (Drummond and Heled, 2010), and phylogeographical models (Lemey et al., 2009, 2010). Like in BEAST 1, an analysis is set up using input XML files. For most standard analyses, these files can be easily created using a graphical user interface (BEAUti 2).

The key difference in design philosophy between BEAST 1 and BEAST 2 is a greater emphasis in the latter on extensibility, resulting in a modular program built around a set of core components. This allows third-party developers to implement new methods as packages that can be added without rebuilding or redeploying BEAST 2. Through such packages, BEAST 2 provides a growing collection of new models not available in BEAST 1, such as flexible birth–death tree-priors (Stadler et al., 2013; Gavryushkina et al., 2014; Kühnert et al., 2016) and structured coalescent models (Vaughan et al., 2014; De Maio et al., 2015), as well as updates to existing models, such as StarBEAST 2 (Ogilvie and Drummond, 2016). A list of available models in BEAST 1 and BEAST 2 can be found at http://beast2.org/beast-features/. (Users should bear in mind that BEAST 2 is modular by design, and thus some third-party packages may not be listed.)

This modular design requires the BEAST 2 user to make active modeling choices, and it is no longer possible to simply perform a "default" analysis. This active involvement opens the door for analyses tailored specifically to particular data sets and questions, greatly

increasing the power of the package. However, it also markedly increases the complexity and makes it easier to inadvertently introduce errors or use inappropriate models. This added complexity could also be daunting to novice users and may result in them preferring simpler, but less powerful, software packages. We will now briefly highlight the key steps required from the BEAST 2 user when running a data analysis.

At its core, BEAST 2 estimates rooted phylogenies ($\mathcal{T}$) from genetic sequencing data ($\mathcal{D}$), with branch lengths in units of calendar time (i.e., the phylogenies are time-trees). It concurrently estimates evolutionary parameters (θ), such as the substitution rate, and parameters describing population dynamics (η), such as speciation/extinction or transmission/recovery rates. For inference, BEAST 2 uses a Markov chain Monte Carlo (MCMC) algorithm to sample from the posterior distribution,

$$\Pr[\mathcal{T}, \eta, \theta | \mathcal{D}] = \frac{\Pr[\mathcal{D} | \mathcal{T}, \theta] \Pr[\mathcal{T} | \eta] \Pr[\eta] \Pr[\theta]}{\Pr[\mathcal{D}]} \ . \quad (1)$$

The output of an analysis is a log-file containing a sample of the states ($\mathcal{T}, \eta, \theta$) visited by the MCMC algorithm. After a so-called burn-in phase, each value ($\mathcal{T}, \eta, \theta$) is visited by the chain at a frequency proportional to its posterior probability, so the output of BEAST 2 (after eliminating the burn-in) is a set of samples from the posterior distribution. A recent book (Drummond and Bouckaert, 2015) describes the general theory and design behind BEAST 2.

For the user to carry out a successful and correct analysis, several steps need to be performed carefully to analyze the data and answer the research question of interest. The researcher must specify a multileveled (i.e., hierarchical) model with several interacting components, including: (i) a suitable model describing the evolution of the sequence data on a time-tree, including the substitution and molecular-clock models ($\Pr[\mathcal{D} | \mathcal{T}, \theta]$); (ii) a phylodynamic model describing the growth of the tree over time ($\Pr[\mathcal{T}, \eta]$); and (iii) sensible prior distributions for each of the parameters of the evolutionary models ($\Pr[\theta]$ and $\Pr[\eta]$).

In addition to the model components, the researcher must also specify and fine-tune MCMC operators that propose new states for the model parameters ($\mathcal{T}, \eta, \theta$). By choosing appropriate proposal algorithms, an MCMC analysis is more likely to sample the posterior distribution efficiently. Finally, once the MCMC chain has sampled a sufficient number of states, the researcher must assess whether the chain has converged and recovered a meaningful signal from the data.

Consequently, the user is challenged with a myriad of choices on the road to a successful analysis. Although many potential pitfalls exist, a simple but solid understanding of the theory behind Bayesian phylogenetic inference can help guide new users through an analysis to reach sound conclusions.

"TAMING THE BEAST" FOR THE USER COMMUNITY

In June 2016, we organized a "Taming the BEAST" workshop in Engelberg, Switzerland, aimed at fostering interaction between BEAST 2 users and developers. The workshop was organized by graduate students and postdoctoral researchers in the Computational Evolution group at ETH Zürich (https://www.bsse.ethz.ch/cevo, with generous financial support from ETH Zürich) and was a mix of lectures by invited speakers (A.J.D., T.A.H., O.G.P., T.G.V., and T.S. were invited speakers.) and hands-on tutorials run by the organisers. (J.B.-S., V.B., L.d.P., D.K., C.M., V.M., N.F.M., J.P., D.A.R., and C.Z. organized the tutorial sessions.) Participants had the opportunity to learn how to use BEAST 2 with help from the developers and to discuss questions specific to their research with other experienced scientists. For the developers, such a workshop provides direct feedback from users on ease-of-use, identifying specific issues and discovering the needs and wishes of the community for future software and methods development.

The workshop was met with great enthusiasm from researchers already using or planning to use BEAST 2, ranging from students to established PIs. (Although originally envisioned for graduate students only, many postdoctoral researchers, some lecturers, and a few professors applied for the workshop as well. Due to the limited capacity and resources, out of 75 applications, we selected 36 participants from 14 countries and 28 universities.) The positive feedback from the participants (see Fig. 1), the overwhelming support from the community and the demand for further workshops has provided motivation to initiate a series of "Taming the BEAST" workshops. At the time of writing, a second successful edition of "Taming the Beast" was run on Waiheke island (New Zealand) in February 2017 and a third edition will take place in July 2017 in London. Further editions are planned for 2018 in Switzerland, and for 2019 and 2020 in locations that are yet to be determined. (We secured funding from ETH Zürich to support the workshop series in 2017–2020.) Each workshop is intended as a global event, allowing users and developers from around the world to meet and share knowledge.

To ensure these resources are available to the community, we have set up a website (https://taming-the-beast.github.io/) with the same name as the workshop series to serve as a platform for collating a comprehensive and cohesive set of BEAST 2 tutorials (see Fig. 2). By providing a set of well-curated tutorials, "Taming the BEAST" offers researchers the resources necessary to learn how to perform analyses in BEAST 2. In addition to tutorials provided by the BEAST 2 developers, this resource page also contains all of the materials (lecture slides, tutorials, data, and example outputs) used during the first two "Taming the BEAST" workshops in Switzerland and New Zealand. These materials will be updated and extended for future editions of the workshop. Tutorials are released under
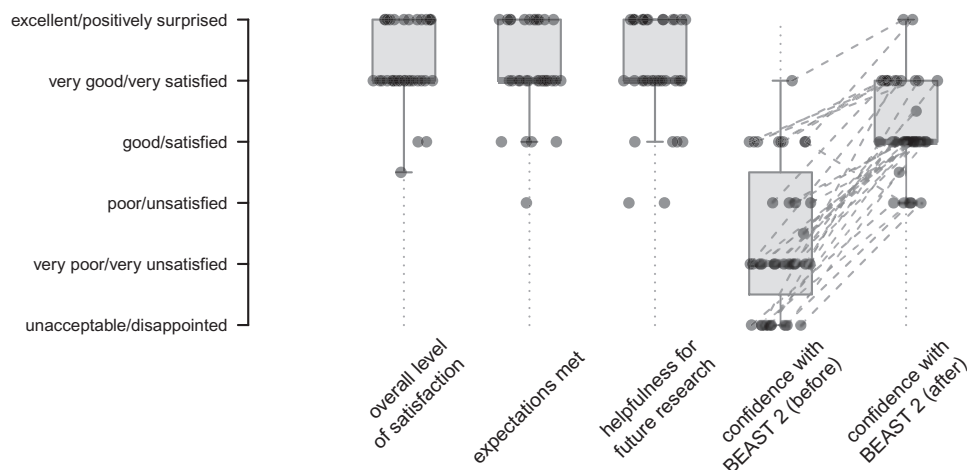
FIGURE 1. Boxplot showing the feedback received from 35 respondents (out of 36 workshop participants) on 5 feedback questions. Of the 35 respondents, all but 3 indicated that they would definitely recommend the workshop to a colleague.
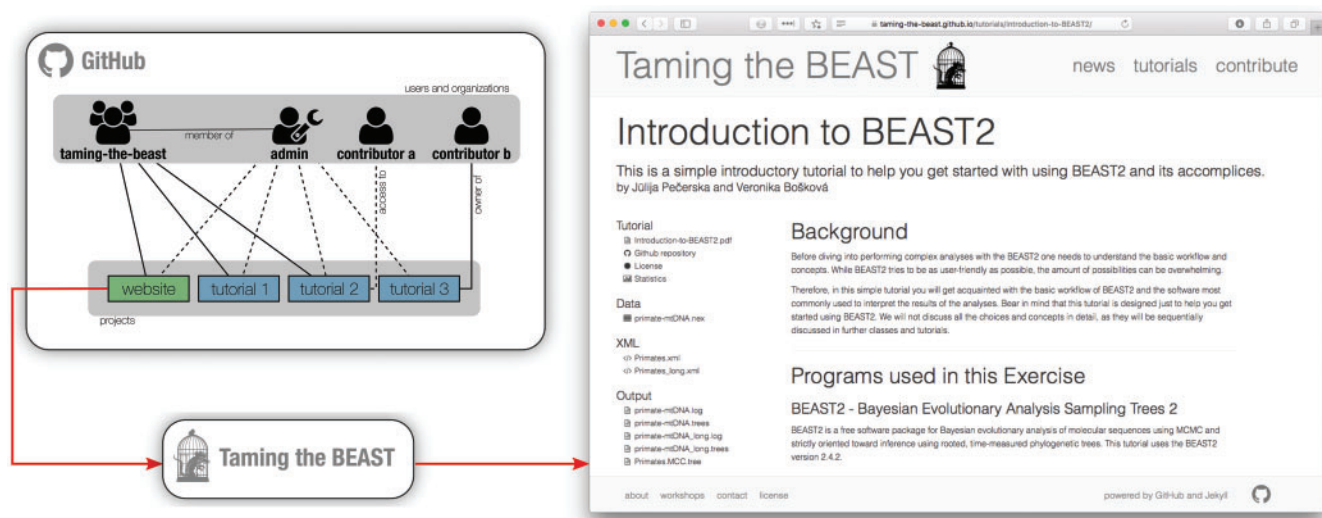

FIGURE 2. Structure of the Taming the BEAST web resource as hosted on GitHub. The diagram on the left shows three possibilities for tutorials available on the website. On the diagram solid lines indicate ownership and dashed lines access. Tutorial 1 is owned by the taming-the-beast organization on GitHub, and does not have any external contributors. Tutorial 2 was created by contributor a, but ownership has been transferred to taming-the-beast. Tutorial 3 was created by contributor b, who has retained ownership. In all three cases, it is essential that at least one of the website administrators has access to the tutorial. The website itself is also hosted on GitHub as a project. When a user visits the website tutorials appear as on the right of the figure. The left panel contains links to a printable PDF version of the tutorial, the data file (or files) used in the tutorial, example BEAST 2 XML files, examples output files and a link to the GitHub repository of the tutorial. Recent changes to the tutorial are also listed.

a license that gives anyone the right to freely use (and modify) tutorials for courses or workshops, as long as appropriate credit is given and the updated material is licensed in the same fashion. (By default we use a Creative Commons Attribution 4.0 license, however the exact license to be used is determined by the tutorial's authors.) We hope that these open resources will encourage other research groups/universities to host and organize their own "Taming the BEAST" workshops. As a community resource, the "Taming the BEAST" website will maintain a list of workshops, and tutorial developers are available to provide support to organizers.

CONTRIBUTING TO TAMING THE BEAST

In keeping with the BEAST 2 design philosophy, we designed the website to have a modular, extensible architecture. Each tutorial is stored in its own GitHub (http://www.github.com) repository, where it is bundled with all of the supporting data and scripts needed to run the tutorial, as well as example output files. This makes it possible for anyone with a GitHub account to raise issues and suggest edits or extensions to tutorials. Similarly, it is also possible for external contributors to submit new tutorials to the website. We provide a template tutorial and comprehensive

documentation to help potential contributors get started.

By providing a "Taming the Beast" platform that allows issues to be raised and content to be edited, we hope that the community will play an active role in curating tutorials. We further envision these resources will continue to grow as the community contributes more tutorials. For instance, the developers of a new BEAST 2 package will be able to add a tutorial for their package to the "Taming the BEAST" site, where it will be accessible in a central location, along with other BEAST 2 tutorials, making it easier for users to become familiar with their package.

Because tutorials are stored in GitHub repositories that track change history, all contributors can receive proper credit for their work. Furthermore, authors of new tutorials can retain ownership of their tutorials after publication. In addition, GitHub tracks traffic to tutorials over time and makes it easy for users to interact with authors, giving authors a measure of their work's impact within the community. Finally, because of the distributed nature of the website, it is robust to changes in any single repository, making it easy to update or add individual tutorials.

Summary

The tutorials on the "Taming the Beast" website allow users to learn about the entire BEAST 2 analysis pipeline, with most tutorials focusing on a particular model component or a single BEAST 2 package. The website provides immediate access to the materials that guide users in the application of a range of models to their own data. In addition, there are tutorials on postprocessing, interpreting results, as well as troubleshooting. We will ensure the maintenance of the website and incorporation of new tutorials through two to three responsible people from the Computational Evolution group at ETH Zürich as well as collaborating groups acting as website administrators. The administrators of the website can be reached via tamingthebeast@bsse.ethz.ch.

We hope that the "Taming the BEAST" platform will allow new BEAST 2 users to accelerate their learning process and to successfully "tame" the BEAST. At the same time, we hope that it will serve as a central repository of teaching materials that will allow BEAST 2 developers and users to exchange knowledge about how to effectively teach the use of BEAST 2. Finally, this platform will hopefully further encourage developers to share their own materials with the wider community.

Acknowledgments

Author's Contributions

J.B.-S., V.B., L.d.P., V.M., and J.P. wrote and submitted the SUK application for starting the "Taming the BEAST" workshop series, with substantial support of C.M. and D.A.R. The first workshop was organized by the whole Computational Evolution group (led by J.B.-S., V.B., and L.d.P.). J.B.-S., V.B., L.d.P., D.K., C.M., V.M., N.F.M., J.P., D.A.R., C.Z., A.J.D., T.A.H., O.G.P., T.G.V., and T.S. wrote the tutorials and/or lecture slides for teaching. L.d.P. created the figures, set up the web resource and GitHub repositories and is the corresponding person regarding these online resources. L.d.P., J.B.-S., V.B., and T.S. wrote the article.

References

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 Comput. Biol. 10(4):e1003537.

De Maio N., Wu C.-H., O'Reilly K.M., Wilson D. 2015. New routes to phylogeography: a Bayesian structured coalescent approximation. PLoS Genet. 11(8):e1005421.

Drummond A.J., Bouckaert R.R. 2015. Bayesian evolutionary analysis with BEAST. Cambridge, UK: Cambridge University Press.

Drummond A.J., Heled J. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27(3):570–580.

Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLOS Biol. 4(5):e88.

Drummond A.J., Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7(1):1.

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(5):1185–1192.

Gavryushkina A., Welch D., Stadler T., Drummond A.J. 2014. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. PLoS Comput. Biol. 10(12):e1003919.

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(4):726–736.

Huelsenbeck J.P., Ronquist F. 2001. Mrbayes: Bayesian inference of phylogenetic trees. Bioinformatics 17(8):754–755.

Kühnert D., Stadler T., Vaughan T.G., Drummond A.J. 2016. Phylodynamics with migration: a computational framework to

quantify population structure from genomic data. Mol. Biol. Evol. 33(8):2102–2116.

Lemey P., Rambaut A., Drummond A.J., Suchard M.A. 2009. Bayesian phylogeography finds its roots. PLOS Comput. Biol. 5(9):e1000520.

Lemey P., Rambaut A., Welch J.J., Suchard M.A. 2010. Phylogeography takes a relaxed random walk in continuous space and time. Mol. Biol. Evol. 27(8):1877–1885.

Ogilvie, H.A., Bouckaert, R.R., Drummond, A.J. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Mol. Biol. Evol. doi: 10.1093/molbev/msx126. [Epub ahead of print].

Stadler T., Kühnert D., Bonhoeffer S., Drummond A.J. 2013. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis c virus (HCV). Proc. Natl. Acad. Sci. USA 110(1): 228–233.

Vaughan T.G., Kühnert D., Popinga A., Welch D., Drummond A.J. 2014. Efficient Bayesian inference under the structured coalescent. Bioinformatics 30(16):2272–9.