babette: BEAUti 2, BEAST2 and Tracer for R

- Richèl J.C. Bilderbeek¹ and Rampal S. Etienne¹
- ¹Groningen Institute for Evolutionary Life Sciences, University of
- Groningen, Groningen, The Netherlands

May 2, 2018

Summary

10

11

12

17

18

19

20

- 1. In the field of phylogenetics, BEAST2 is one of the most widely used software tools. It comes with the graphical user interfaces BEAUti 2, DensiTree and Tracer, to create BEAST2 configuration files and to interpret BEAST2's output files. However, when many different alignments or model setups are required, a workflow of graphical user interfaces is cumbersome.
- 2. Here, we present a free, libre and open-source package, babette:
 'BEAUti 2, BEAST2 and Tracer for R', for the R programming language.
 babette creates BEAST2 input files, runs BEAST2 and parses its results, all from an R function call.
- 3. We describe babette's usage and the novel functionality it provides compared to the original tools and we give some examples.
- 4. As babette is designed to be of high quality and extendable, we conclude by describing the further development of the package.

Keywords: computational biology, evolution, phylogenetics, BEAST2, R

3 1 Introduction

Phylogenies are commonly used to explore evolutionary hypotheses. Not only can phylogenies show us how species (or other evolutionary units) are related to each other, but we also estimate relevant parameters such as extinction and speciation rates. There are many phylogenetics tools available to obtain an esti-27 mate of the phylogeny of a given set of species. BEAST2 (Bouckaert et al. 2014) is one of the most widely used ones. It uses a Bayesian statistical framework to estimate the joint posterior distribution of estimated phylogenies and model parameters, from one or more DNA, RNA or amino acid alignments (see figure 31 1 for an overview of the workflow). It has a graphical and a command-line interface, that both need a config-33 uration file containing alignments and model parameters. BEAST2 is bundled with BEAUti 2 (Drummond et al. 2012) ('BEAUti' from now on), a desktop application to create a BEAST2 configuration file. BEAUti has a user-friendly 36 graphical user interface, with helpful default settings. As such, BEAUti is an attractive alternative to manual and error-prone editing of BEAST2 configuration files. However, BEAUti cannot be called from a command-line script. This implies that when the user wants to explore the consequences of various settings, this must be done manually. This is the managable workflow when using a few alignments and doing a superficial analysis of sensitivity of the reconstructed tree to model settings. For exploring many trees (for instance from simulations) and for more thorough sensitivity analysis, one would like to loop through mul-45 tiple (simulated) alignments, nucleotide substitution models, clock models and tree priors. One such tool to replace BEAUti is BEASTmasteR (Matzke 2015), 47 which focuses on morphological traits and tip-dating, but also supports DNA data. BEASTmasteR, however, requires hundreds of lines of R code to setup the ₅₀ BEAST2 model configuration and a Microsoft Excel file to specify alignment files.

BEAST2 is also associated with Tracer (Rambaut & Drummond 2007) and
DensiTree (Bouckaert & Heled 2014). Both are desktop applications to analyze the output of BEAST2, each with a user-friendly graphical user interface. Tracer's purpose is to analyze the parameter estimates generated from a
BEAST2 run. It shows, among others, the effective sample size (ESS) and time
series ('the trace', hence the name) of each variable in the MCMC run. Both
ESS and trace are needed to assess the strength of the inference. DensiTree visualizes the phylogenies of a BEAST2 posterior, with many options to improve
the simultaneous display of many phylogenies.

However, for exploring the output of many BEAST2 runs, one would like a script to collect all parameters' ESSes, parameter traces and posterior phylogenies. There is no single package that offers a complete solution, but examples of R packages that offer a partial solution are rBEAST (Ratmann 2015) and RBeast (Faria & Suchard 2015). RBeast provides some plotting options and parsing of BEAST2 output files, but the plotting functions are too specific for general use, whilst the functions for parsing call those of babette. rBEAST was developed to test a particular biological hypothesis (Ratmann et al. 2016), and hence was not designed for general use.

Here, we present babette: BEAUti 2, BEAST2 and Tracer for R, which
creates BEAST2 (v.2.4.7) configuration files, runs BEAST2, and analyzes its
results, all from an R function call. This will save time, tedious mouse clicking
and reduces the chances of errors in such repetitive actions. The interface of
babette mimics the tools it is based on. This familiarity helps both beginner
and experienced BEAST2 users to make the step from those tools to babette.
babette enables the creation of a single-script pipeline from sequence alignments

$_{\scriptscriptstyle{78}}$ 2 Description

babette is written in the R programming language (R Core Team 2013) and enables the full BEAST2 workflow from an R function call, in a similar way 80 to what BEAUti, DensiTree and Tracer do. babette's main function is run, which configures BEAST2, runs it and parses its output. run needs at least the name of a FASTA file containing a DNA alignment. The default settings for the other arguments of run are identical to BEAUti's and BEAST2's default 84 settings. Per alignment, a site model, clock model and tree prior can be chosen. Multiple alignments can be used, each with its own (unlinked) site model, clock model and tree prior. babette currently has 108 exported functions to set up a BEAST2 configuration file. babette can currently handle a majority of BEAUti use cases. Because of BEAUti's high number of plugins, babette uses a software architecture that is designed to be extended. Furthermore, babette has 13 exported functions to run and help run BEAST2. One function is used to run BEAST2, another one installs BEAST2 to a default location. Finally, babette has 21 exported function to parse the BEAST2 output files and analyze the created posterior. babette gives the same ESSes and summary statistics as Tracer. The data is formatted such that it can easily be visualized using ggplot2 (for a trace, similar to Tracer) or phangorn (Schliep 2011) (for the phylogenies in a posterior, similar to DensiTree). Currently, babette does not replace all functionality in BEAUti, BEAST2 and their many plug-ins. This is reasonable, as these themselves also change 100 in time. babette provides the majority of default tree priors and works only 101 on DNA data. babette supports the most important command-line arguments 102

of BEAST2, provides the core Tracer analysis options, and has the most basic subset of plotting options of DensiTree. Up until now, the babette features implemented are those requested by users. Further extension of babette will be based on future user requests.

$_{ ext{\tiny 107}}$ 3 Usage

```
babette can be installed easily with the devtools package (Wickham & Chang
108
   2016):
    library(devtools)
110
    install_github("richelbilderbeek/babette")
111
   To start using babette, load its functions in the global namespace first:
112
    library(babette)
   For most convenience, let babette install BEAST2 to a default location:
114
    install_beast2()
   BEAUti, and likewise babette, needs at least a FASTA filename to produce a
116
   BEAST2 configuration file. In BEAUti, this is achieved by loading a FASTA file,
117
   then saving an output file using a common save file dialog. After this, BEAST2
   needs to be applied to the created configuration file. It creates multiple files
119
   storing the posterior. These output files must be parsed by either Tracer or
   DensiTree. In babette, all this is achieved by:
121
    out <- run(fasta_filenames = "anthus_aco.fas")</pre>
122
    This code will create a (temporary) BEAST2 configuration file, from the FASTA
123
   file with name anthus_aco.fas (which is supplied with the package, from
    (Van Els & Norambuena 2018)), using the same default settings as BEAUti,
125
    which are, among others, a Jukes-Cantor site model, a strict clock, and a Yule
```

```
birth tree prior. babette will then execute BEAST2 using that file, and parses
   the output. The returned data structure, named out, is a list of parameter
128
   estimates (called estimates), posterior phylogenies (called anthus_aco_trees,
129
   named after the alignment's name) and MCMC operator performance (operators).
130
   An example of using a different site model, clock model and tree prior is:
131
    out <- run(
132
      fasta_filenames = "anthus_aco.fas",
133
      site_models = create_hky_site_model(),
134
      clock_models = create_rln_clock_model(),
135
      tree_priors = create_bd_tree_prior()
136
   )
137
   This code uses an HKY site model, a relaxed log-normal clock model and a birth-
138
   death tree prior, each with their default settings in BEAUti. Table 1 shows an
139
   overview of all functions to create site models, clock models and tree priors. Note
140
   that the arguments' names site_models, clock_models and tree_priors are
   plural, as each of these can be (a list of) one or more elements. Each of these
142
   arguments must have the same number of elements, so that each alignment has
   its own site model, clock model and tree prior. An example of two alignments,
144
   each with its own site model, is:
    out <- run(
146
      fasta_filenames = c(
147
         "anthus_aco.fas",
         "anthus_nd2.fas"
      ),
150
      site_models = list(
151
         create_tn93_site_model(),
152
         create_gtr_site_model()
153
      )
154
```

```
)
155
    babette also uses the same default prior distributions as BEAUti for each of
156
    the site models, clock models and tree priors. For example, by default, a Yule
157
    tree prior assumes that the birth rate follows a uniform distribution, from minus
158
    infinity to plus infinity. One may prefer a different ddistribution instead. Here
    is an example how to specify an exponential distribution for the birth rate in a
160
    Yule tree prior in babette:
    out <- run(
162
163
      fasta_filenames = "anthus_aco.fas",
      tree_priors = create_yule_tree_prior(
         birth_rate_distr = create_exp_distr()
165
      )
166
    )
167
    In this same example, one may specify the initial shape parameters of the expo-
168
    nential distribution. In BEAST2's implementation, an exponential distribution
    has one shape parameter: its mean, which can be set to any value with BEAUti.
170
    Within babette, to set the mean value of the exponential distribution to a fixed
171
    (non-estimated) value, do:
172
    out <- run(
173
      fasta_filenames = "anthus_aco.fas",
174
      tree_priors = create_yule_tree_prior(
175
         birth_rate_distr = create_exp_distr(
176
           mean = create_mean_param(
177
              value = 1.0,
178
              estimate = FALSE
179
           )
180
         )
      )
182
```

```
NOTE: suggest to remove this paragraph until node dating Our initial
184
   motivation to create babette was that we wanted to fix the crown age (the
    time of first divergence) of a phylogeny. BEAUti assumes that a phylogeny has
186
    a crown age that needs to be jointly estimated with the phylogeny and other
   parameters. It does not allow for fixing the crown age. Without babette,
188
    one needs to manually edit the BEAST2 configuration file, to remove all tree
   operators that change the height of the root node, which is tedious and prone
190
   to errors. Fixing the crown ages is useful when it is unnecessary to accurately
191
   model the uncertainty of absolute divergence times. This is how to specify a
192
   fixed crown age with babette:
193
    out <- run(
194
      fasta_filenames = "anthus_aco.fas",
195
      posterior_crown_age = 15
196
   )
197
   babette also supports node dating. Like BEAUti, one can specify Most Recent
    Common Ancestor ('MRCA') priors. An MRCA prior allows to specify taxa
199
   having a common ancestor, including a distribution for the date of that ancestor.
    An application of such a prior is to specify a (close to) fixed crown age, by
201
   including all taxa and assuming a tight normal distribution around the desired
202
   crown age. Within babette, this is achieved as such:
203
    posterior <- run(
      fasta_filenames = "anthus_aco.fas",
205
      mrca_priors = create_mrca_prior(
206
         taxa_names = get_taxa_names("anthus_aco.fas"),
207
         alignment_id = get_alignment_id("anthus_aco.fas"),
208
```

183)

209

mrca_distr = create_normal_distr(

```
mean = create_mean_param(value = 15.0, estimate =
                FALSE),
211
           sigma = create_sigma_param(value = 0.01, estimate =
212
                FALSE)
213
         )
214
      )
215
    )
216
    babette allows for the same core functionality as Tracer, in which it works on
217
    the values of the parameter estimates sampled in the BEAST2 run. This is
218
    called the "trace" (hence the name). The start of the trace is usually discarded,
    as an MCMC algorithm (such as used by BEAST2) first has to converge to
220
    its equilibrium. The start of the trace, called the "burn-in", will be removed,
221
    because its parameter estimates are not representative. By default, Tracer dis-
222
    cards the first 10% of all the parameter estimates. To remove a 20% burn-in
223
    from all parameter estimates in babette, the following code can be used:
224
    traces <- remove_burn_ins(</pre>
225
      traces = out$estimates,
226
      burn_in_fraction = 0.2
227
    )
228
    Tracer shows the ESSes of each posterior's variables. These ESSes are important
229
    to determine the strength of the inference. As a rule of thumb, an ESS of 200 is
    acceptable for any parameter estimate. To calculate the effective sample sizes
231
    (of all estimated variables) in babette:
    esses <- calc_esses(
233
      traces = traces,
234
      sample_interval = 1000
   )
236
```

210

```
Tracer displays multiple summary statistics for each estimated variable: the
    mean and its standard error, standard deviation, variance, median, mode, geo-
    metric mean, 95% highest posterior density interval, auto-correlation time and
239
    effective sample size. It displays these statistics per variable. In babette, these
    summary statistics are collected for all estimated parameters at once:
241
    sum_stats <- calc_summary_stats(</pre>
242
      traces = traces,
243
      sample_interval = 1000
244
245
    babette allows for the same functionality as DensiTree. DensiTree displays the
    phylogenies in a posterior at the same time scale, drawn one over one another,
247
    allowing to see the uncertainty in topology and branch lengths. Within the
    object out, the posterior phylogenies are stored as anthus_aco_trees, and can
249
    be plotted as such:
250
    plot_densitree(phylos = out$anthus_aco_trees)
251
    Instead of running the full pipeline, babette also allows to only create a BEAST2
    configuration file. To create a BEAST2 configuration file, with all settings to
253
    default, use:
254
    create_beast2_input_file(
      input_filenames = babette::get_babette_path("anthus_aco.
           fas"),
      output_filename = "beast2.xml"
258
    )
259
    This file can then be loaded and edited by BEAUti, or run by BEAST2. babette
260
    also allows to only run BEAST2 from a BEAST2 configuration file:
    run_beast2(
      input_filename = "beast2.xml",
263
```

```
output_log_filename = "run.log",

output_trees_filenames = "posterior.trees",

output_state_filename = "final.xml.state"

in this example, we specify the names of the desired BEAST2 output files. These

output files can then be inspected with other tools, or to be used to continue a
```

270 BEAST2 run. Also run supports to specify these filenames, and these files are
271 put in a temporary folder by default.

272 4 babette resources

babette is free, libre and open source software available at http://github. com/richelbilderbeek/babette and is licensed under the GNU General Pub-274 lic License v3.0. babette uses the Travis CI (https://travis-ci.org) con-275 tinuous integration service, which is known to significantly increase the number 276 of bugs exposed (Vasilescu et al. 2015) and increases the speed at which new 277 features are added (Vasilescu et al. 2015). babette has a 100% code cover-278 age, which correlates with code quality (Horgan et al. 1994; Del Frate et al. 279 1995). babette follows Hadley Wickham's style guide (Wickham 2015), which improves software quality (Fang 2001). babette depends on multiple packages, 281 which are ape (Paradis et al. 2004), beautier (Bilderbeek 2018b), beastier (Bilderbeek 2018a), devtools (Wickham & Chang 2016), geiger (Harmon et al. 283 2008), ggplot2 (Wickham 2009), knitr (Xie 2017), phangorn (Schliep 2011), rmarkdown (Allaire et al. 2017), seqinr (Charif & Lobry 2007), stringr (Wick-285 ham 2017), testit (Xie 2014) and tracerer (Bilderbeek 2018c). We tested 286 babette to give a clean error message for incorrect input, by calling babette 287 one million times with random or random sensible inputs, using the Peregrine 288 high performance computer cluster. The scripts to do so are supplied with

babette. babette's development takes place on GitHub, https://github.com/richelbilderbeek/ babette, which accommodates collaboration (Perez-Riverol et al. 2016) and im-292 proves transparency (Gorgolewski & Poldrack 2016). babette's GitHub facili-293 tates feature requests and has guidelines how to do so. 294 babette's documentation is extensive. All functions are documented in the package's internal documentation. For quick use, each exported function shows a minimal example. For easy exploration, each exported function's documen-297 tation links to related functions. Additionally, babette has a vignette that demonstrates extensively how to use it. There is documentation on the GitHub 299 to get started, with a dozen examples of BEAUti screenshots with equivalent babette code. Finally, babette has tutorial videos that can be downloaded or 301

5 Citation of babette

viewed on YouTube, https://goo.gl/weKaaU.

Scientists using babette in a published paper can cite this article, and/or cite
the babette package directly. To obtain this citation from within an R script,
use:

> citation("babette")

$_{ iny 808}$ 6 Acknowledgements

Thanks to Yacine Ben Chehida and Paul van Els for supplying their BEAST2 use cases. Thanks again to Paul van Els for sharing his FASTA files for use by this package. Thanks to Leonel Herrera-Alsina, Raphael Scherrer and Giovanni Laudanno for their comments on this package and article. Thanks to Huw Ogilvie and one anonymous reviewer for reviewing this article. Thanks to

- ³¹⁴ rOpenSci, and especially Noam Ross, Guangchuang Yu, David Winter and Jolle
- Barido-Sottani for reviewing the package's source code. We would like to thank
- the Center for Information Technology of the University of Groningen for their
- support and for providing access to the Peregrine high performance computing
- 318 cluster.

³¹⁹ 7 Authors' contributions

- RJCB and RSE conceived the idea for the package. RJCB created and tested
- the package, and wrote the first draft of the manuscript. RSE contributed
- 322 substantially to revisions.

323 References

- Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wick-
- ham, H., Cheng, J. & Chang, W. (2017) rmarkdown: Dynamic Documents
- for R. R package version 1.8.
- Bilderbeek, R.J. (2018a) beastier: BEAST2 from R. https://github.com/
- richelbilderbeek/beastier [Accessed: 2018-03-16].
- Bilderbeek, R.J. (2018b) beautier: BEAUti 2 from R. https://github.com/
- richelbilderbeek/beautier [Accessed: 2018-03-16].
- Bilderbeek, R.J. (2018c) tracerer: Tracer from R. https://github.com/
- richelbilderbeek/tracerer [Accessed: 2018-03-16].
- Bouckaert, R. & Heled, J. (2014) Densitree 2: Seeing trees through the forest.
- bioRxiv, p. 012401.
- 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**, e1003537.
- ³³⁸ Charif, D. & Lobry, J. (2007) SeqinR 1.0-2: a contributed package to the R
- project for statistical computing devoted to biological sequences retrieval and
- analysis. U. Bastolla, M. Porto, H. Roman & M. Vendruscolo, eds., Struc-
- tural approaches to sequence evolution: Molecules, networks, populations, Bi-
- ological and Medical Physics, Biomedical Engineering, pp. 207–232. Springer
- ³⁴³ Verlag, New York. ISBN: 978-3-540-35305-8.
- Del Frate, F., Garg, P., Mathur, A.P. & Pasquini, A. (1995) On the correlation
- between code coverage and software reliability. Software Reliability Engi-
- neering, 1995. Proceedings., Sixth International Symposium on, pp. 124–132.
- 347 IEEE.
- Drummond, A.J., Suchard, M.A., Xie, D. & Rambaut, A. (2012) Bayesian phy-
- logenetics with beauti and the beast 1.7. Molecular biology and evolution, 29,
- 1969–1973.
- Fang, X. (2001) Using a coding standard to improve program quality. Quality
- Software, 2001. Proceedings. Second Asia-Pacific Conference on, pp. 73–78.
- 353 IEEE.
- Faria, N. & Suchard, M.A. (2015) RBeast. https://github.com/beast-dev/
- 355 RBeast [Accessed: 2018-03-02].
- 356 Gorgolewski, K.J. & Poldrack, R. (2016) A practical guide for improving trans-
- parency and reproducibility in neuroimaging research. bioRxiv, p. 039354.
- Harmon, L., Weir, J., Brock, C., Glor, R. & Challenger, W. (2008) Geiger:
- investigating evolutionary radiations. *Bioinformatics*, **24**, 129–131.

- Horgan, J.R., London, S. & Lyu, M.R. (1994) Achieving software quality with testing coverage measures. *Computer*, **27**, 60–69.
- Matzke, N.J. (2015) BEASTmasteR: R tools for automated conversion of
- NEXUS data to BEAST2 XML format, for fossil tip-dating and other uses.
- https://github.com/nmatzke/BEASTmasteR [Accessed: 2018-02-28].
- Paradis, E., Claude, J. & Strimmer, K. (2004) APE: analyses of phylogenetics
- and evolution in R language. *Bioinformatics*, **20**, 289–290.
- Perez-Riverol, Y., Gatto, L., Wang, R., Sachsenberg, T., Uszkoreit, J., Lepre-
- vost, F., Fufezan, C., Ternent, T., Eglen, S.J., Katz, D.S. et al. (2016) Ten
- simple rules for taking advantage of git and github. bioRxiv, p. 048744.
- R Core Team (2013) R: A Language and Environment for Statistical Computing.
- R Foundation for Statistical Computing, Vienna, Austria.
- Rambaut, A. & Drummond, A.J. (2007) Tracer v1.4. Available from
- http://beast.bio.ed.ac.uk/Tracer.
- Ratmann, O. (2015) rBEAST. https://github.com/olli0601/rBEAST [Ac-
- cessed: 2018-03-02].
- Ratmann, O., Van Sighem, A., Bezemer, D., Gavryushkina, A., Jurriaans, S.,
- Wensing, A., De Wolf, F., Reiss, P., Fraser, C. et al. (2016) Sources of hiv
- infection among men having sex with men and implications for prevention.
- Science translational medicine, 8, 320ra2-320ra2.
- Schliep, K. (2011) phangorn: phylogenetic analysis in R. Bioinformatics, 27,
- ³⁸¹ 592–593.
- Van Els, P. & Norambuena, H.V. (2018) A revision of species limits in neotrop-
- ical pipits anthus based on multilocus genetic and vocal data. *Ibis*.

- ³⁸⁴ Vasilescu, B., Yu, Y., Wang, H., Devanbu, P. & Filkov, V. (2015) Quality and
- productivity outcomes relating to continuous integration in github. Proceed-
- ings of the 2015 10th Joint Meeting on Foundations of Software Engineering,
- pp. 805–816. ACM.
- Wickham, H. (2009) ggplot2: elegant graphics for data analysis. Springer New
- York.
- Wickham, H. (2015) R packages: organize, test, document, and share your code.
- O'Reilly Media, Inc.
- Wickham, H. (2017) stringr: Simple, Consistent Wrappers for Common String
- Operations. R package version 1.2.0.
- Wickham, H. & Chang, W. (2016) devtools: Tools to Make Developing R Pack-
- ages Easier. R package version 1.12.0.9000.
- 396 Xie, Y. (2014) testit: A Simple Package for Testing R Packages. R package
- version 0.4, http://CRAN.R-project.org/package=testit.
- 398 Xie, Y. (2017) knitr: A General-Purpose Package for Dynamic Report Genera-
- tion in R. R package version 1.17.

Name	Description
run	Run BEAST2
create_gtr_site_model	Create a GTR site model
create_hky_site_model	Create an HKY site model
create_jc69_site_model	Create a Jukes-Cantor site model
create_tn93_site_model	Create a TN93 site model
create_rln_clock_model	Create a relaxed log-normal clock model
create_strict_clock_model	Create a strict clock model
create_bd_tree_prior	Create a birth-death tree prior
create_cbs_tree_prior	Create a coalescent Bayesian skyline tree prior
create_ccp_tree_prior	Create a coalescent constant-population tree prior
create_cep_tree_prior	Create a coalescent exponential-population tree prior
create_yule_tree_prior	Create a Yule tree prior
create_beta_distr	Create a beta distribution
create_exp_distr	Create an exponential distribution
create_gamma_distr	Create a gamma distribution
create_inv_gamma_distr	Create an inverse gamma distribution
create_laplace_distr	Create a Laplace distribution
create_log_normal_distr	Create a log-normal distribution
create_normal_distr	Create a normal distribution
create_one_div_x_distr	Create a 1/X distribution
create_poisson_distr	Create a Poisson distribution
create_uniform_distr	Create a uniform distribution

Table 1: babette's main functions