tbea: tools for pre- and post-processing in Bayesian evolutionary analyses

Gustavo A. Ballen[1,2,*] and Sandra Reinales[3]

[1] Instituto de Biociências de Botucatu, Universidade Estadual Paulista "Júlio de Mesquita Filho", Botucatu, SP, Brazil. Orcid https://orcid.org/0000-0001-5424-8608

[2] School of Biological and Behavioural Sciences, Queen Mary University of London, London, United Kingdom

[3] Departamento de Botânica, Instituto de Biociências, Universidade de São Paulo, SP, Brazil. Orcid https://orcid.org/0000-0003-3071-9190

∗ Correspondence: gustavo.a.ballen@gmail.com, gustavo.ballen@unesp.br

1 Abstract

1. Estimating phylogenies in which branch lengths are expressed in units of absolute time is crucial for testing hypotheses in modern evolutionary biology. However, bioinformatic tools to process data from bayesian divergence time estimation analyses are often not easily adaptable, and documenting methodological choices is not a generalized practice.

2. The R package tbea is a toolset to integrate biological, geological and paleontological information to optimize the specification of models, their parameters and prior distributions in divergence times estimation analyses. tbea implements statistical models to (i) better translate time information in dating sources into the specified calibration densities, (ii) improve comparisons between prior and posterior distributions for parameters of interest, (iii) carry out inference on origination times for a collection of posterior distributions, (iv) summarize different distributions into a single one, useful e.g. when building secondary calibrations, and (v) improve the reproducibility of divergence time estimation analyses allowing users to document methodological choices.

3. We illustrate the package functionalities by carrying out two worked examples. One on the phylogenetic relationships and divergence time estimation of South American Saber-Tooth Characins of the family Cynodontidae, and another one on the separation time of drainages East and West of the Andes in South America during the Neogene.

4. It is expected that the tools herein available will be key when estimating events in time from collections of point estimates, as well as the combination of different posterior densities from the same parameter are useful to justifying the selection of secondary calibration points, or discussing the timing of biogeographic events when multiple sources are available.

Keywords: biogeography, computational biology, data visualization, divergence time estimation, fossil record, morphology, reproducibility, secondary and tip calibration, total evidence analysis

2 Introduction

In a phylogenetic tree, the characters, either molecular or morphological, only contain information about how closely or distantly related the terminals are. Branch lengths represent a combination of the evolutionary rate between and the time required to accumulate such a change between nodes. The evolutionary rate and time are intrinsically linked and statistically non-identifiable in standard tree inference; hence, it is not possible to distinguish between higher evolutionary rates over a short time interval and slow evolutionary rates over a long time. Therefore, estimating a phylogeny in which branch lengths are in units of absolute time (i.e. time tree or chronogram) is a critical step in many comparative analyses such as modeling morphological evolution, testing biogeographic hypotheses, and estimating lineage diversification rates. One key reason why time trees are useful lies in the nature of evolutionary models which consider a stochastic process in time, such as the discrete Markov models used in both morphological and molecular evolution (Lewis, 2001; Yang, 2014). Therefore, branch length units in expected number of changes per site, such as those estimated in pure phylogenetic analysis (e.g., ML or Bayesian inference) are inadequate for such analyses and therefore divergence time estimation for separating evolutionary rate and time have demonstrated to be key in evolutionary biology (Warnock and Wright, 2021).

Independent temporal evidence is required to calibrate phylogenies to absolute time. This information can come from a variety of sources, including geological date estimates from fossils or biogeographic events (primary calibrations; Heath et al., 2014; Landis, 2017), time estimates obtained from previous studies (secondary calibration), or dated tip sequences from serially sampled data (typically time-stamped pathogen data; Stadler and Yang, 2013). One of the most common strategies for using independent time data to calibrate a phylogeny is assigning time constraints to one or more nodes in a phylogeny, usually called node calibrations (Rannala and Yang, 2007). Nevertheless, when using fossil information, it is possible to account that the fossils are part of the same diversification process that generated the extant species and incorporate them as dated tips alongside living relatives (tip calibration; Ronquist et al., 2012).

Several methods are available to estimate time trees using one of the above-mentioned approaches and can be grouped into three main types: Methods based on distances, maximum likelihood, and Bayesian inference. The first group includes methods such as least-squares dating (LSD; (To et al., 2015)), whereas the second group includes penalized likelihood (Sanderson, 2002; Smith and O'Meara, 2012). These two groups are frequently applied to analyze phylogenomic datasets containing hundreds to thousands of sequences because of their computational efficiency (Barba-Montoya et al., 2021). Lastly, Bayesian-based methods, such as those based on the birth-death process (Stadler, 2010; Heath et al., 2014), and methods based on the multispecies coalescent process (Heled and Drummond, 2009; Rannala and Yang, 2017), consider uncertainty in fossil calibrations and incorporate into the models previous knowledge (or lack thereof) of the amount of elapsed time from the ancestral node to its calibrating fossil using prior distributions.

In a very general form, we can describe a Bayesian divergence time estimation model as

$$P(\Theta|X,C) = \frac{P(X|\theta,\psi_s)P(C|\psi_t)P(\psi_s|\psi_t\theta)P(\psi_t,\theta)}{P(X,C)} \tag{1}$$

where $P(\Theta|X,C)$ is the posterior distribution of the model parameters, $P(X|\theta,\psi_s)$ is the phylogenetic likelihood of the sequence data X, $P(C|\psi_t)$ are the (node) time calibration densities, $P(\psi_s|\psi_t\theta)$ is the clock model, $P(\psi_t,\theta)$ is the tree prior, and $P(X,C)$ is a normalizing constant.

Divergence time estimation proceeds by taking sequence data, a set of distributions to be used as time calibrations, and the choice of appropriate models for the clock and the tree prior. The phylogenetic likelihood is pretty standard and is also found in phylogenetic maximum likelihood inference. The sequence substitution model is herein considered as part of the phylogenetic likelihood expression. The method then estimates branch lengths, and uses all the clock model, tree prior, and calibration distributions in order to calculate the posterior distribution of all the parameters in the model.

Since Equation 1 is analytically intractable, MCMC sampling algorithms are used to approximate the posterior distribution numerically.

The result is then a time tree distribution, in which the branch lengths are in units of absolute time (usually Ma). Since node times (and branch lengths) are also parameters in the tree ψ, the MCMC sampler returns a posterior distribution for each node time. When the tree ψ is fixed, we only have this parameter set and their uncertainties, but when the tree is coestimated with the divergence times, we also have a posterior distribution of trees, which may or may not share nodes across the different elements of the distribution. In the end, a Bayesian analysis of any sort aims to characterize the posterior distributions and thus the uncertainty on estimated parameters rather than point estimates, which are the main goal of maximum likelihood.

One of the challenges in node-calibration strategies is to translate the inherent uncertainty in fossil dating into a prior distribution. Some programs such as `Beast2` (Bouckaert et al., 2019) require the users to input parameter values for the node calibration densities, often by eye with the aid of plots as is done in `beauti`. On the other hand, when different time estimates for the same event (e.g. a node) are available, summarizing such information to constrain a node age is not straightforward. Consequently, it is common to see in the literature approaches which rely on manual specification to find a single and representative distribution, or all sorts of arguments to choose among alternatives. Nevertheless, a reproducible method to summarize different sources of information is desirable, because every study constitutes evidence, as long as the estimates are independent.

In addition, bioinformatic tools to pre- and post-process data from Bayesian divergence time estimation analyses are often limited to specific tools, and highly idiosyncratic. Then, it is common to be forced to use several programs and packages to prepare input data (e.g. combining morphological and molecular data), and analyze the results, especially when comparing prior and posterior node age densities is a goal.

The package `tbea` has tools for improving the reproducibility of the generation of concatenated alignments X, express better the time information in calibration sources onto the specified calibrations $P(C|\psi_t)$, propose adequate initial values for the time tree ψ_t, improve comparisons between prior and posterior distributions for parameters of interest, in particular node times, and carry out inference on origination times for a collection of posterior distributions of the same class in different studies. Finally, the package includes tools for summarizing tree distributions in terms of both topology and branch lengths. It also has two methods for estimating the time of an event where independent estimates ($P_{i:N}(\Theta_i|X_i, C_i)$) are available (e.g., multiple estimates of the timing of a biogeographic event) and finally implements conflation of densities as a tool for summarizing multiple distributions of the same parameter into a single one.

3 Materials and methods

3.1 Implementation

The `tbea` package is written in pure R (R Core Team, 2024) without the need of compiled code, simplifying its installation process. One important principle used for the development of the package is to rely on tools from the base package of R as much as possible, as well as reducing to the minimum possible the package dependencies. The reason for this is to minimize the chance of breaking code due to breaking changes in dependencies when the number of packages used is large, or with complicated development such as packages form the `tidyverse` (Wickham et al., 2019). It is expected that code relying mainly on the base R package will enjoy the high level of back-compatibility available in R, therefore improving the reproducibility of code using `tbea`. The package follows modern development techniques (Wickham, 2015) including thorough code documentation, unit testing, and version control. The code is open source for both the stable release on CRAN `https://cran.r-project.org/package=tbea`, and development versions on github `https://github.com/gaballench/tbea`.

3.2 Working examples

We introduce the package and illustrate its functionalities by carrying out two worked examples. One on the phylogenetic relationships and divergence time estimation of South American Sabre-Tooth Characins of the family Cynodontidae, and another one on the separation time of drainages East and West of the Andes in South America. The former targets pre-analysis tools and a couple of functionalities for post-analysis settings, whereas the latter aims at demonstrating the use of post-analysis tools when information from different analysis is available and the goal is to infer a single parameter, in this case, the separation time between these two biogeographic regions.

In all the code examples below it is assumed that the package was loaded already in the R session with `library(tbea)`. Whenever other packages are necessary these are loaded at the beginning of the appropriate code block.

3.2.1 The South American Sabertooth Characins

We will use the South American Sabre-Tooth Characins of the family Cynodontidae for a divergence time estimation analysis. The phylogenetic relationships of the group were revisited by Ballen et al. (2022), although these authors did not provide information about the temporal scale of their evolutionary relationships. We will use the morphological and molecular datasets of that study, including a fossil terminal.

The dataset comprises 10 terminals, eight molecular markers, and one morphological partition with 80 characters. We refer the reader to the study for details on the substitution models and other details of the phylogenetic analysis.

Divergence time estimation was carried out using `Beast2` (Bouckaert et al., 2019) with the `SampledAncestors` package (Gavryushkina et al., 2014). Preparation of the initial tree was carried out using the package `ape` (Paradis and Schliep, 2019). Post-analysis tree plotting was carried out in R (R Core Team, 2024).

3.2.2 The separation of drainages East and West of the Andes

Divergence time estimation data were compiled from the literature that reported separation events between cis- and trans-Andean lineages, be it contemporary species or nodal divergences (Abe et al., 2014; Cheviron et al., 2005; Collins and Dubach, 2000; Cortés-Ortiz et al., 2003; D'Horta et al., 2013; Dick et al., 2003; Elias et al., 2009; Fernandes et al., 2014; Grau et al., 2005; Gutiérrez et al., 2014; Hardman and Lundberg, 2006; Hernández Torres, 2015; Machado et al., 2014; Miller et al., 2008; Patané et al., 2009; Patel et al., 2011; Picq et al., 2014; Ribas et al., 2005, 2007; Říčan et al., 2013; Ruiz-García et al., 2015; Smith et al., 2014; Voss et al., 2013; Weir and Price, 2011). A total of 54 data points were compiled along with their credible intervals, most frequently HPD intervals.

The distribution of divergence time data is very asymmetrical with a heavy right tail, product of several outliers older than 10 Ma (Figure 1). These outliers affect strongly those methods who make a stronger assumption on the lack of uncertainty in the measurement of time values (e.g., methods based on stratigraphic intervals and less strongly, those based on cumulative densities). Given this analytical issue, these values must be excluded, what makes unnecessary the process of linearization through data transformation given that the data conform to a linear model after outlier remotion.

4 Results

Below we demonstrate the different contexts of divergence time estimation studies in which the `tbea` package can be applied. First a short overview of the goal is presented, and then a code sample illustrating how to reach it, finally making any remarks when necessary.

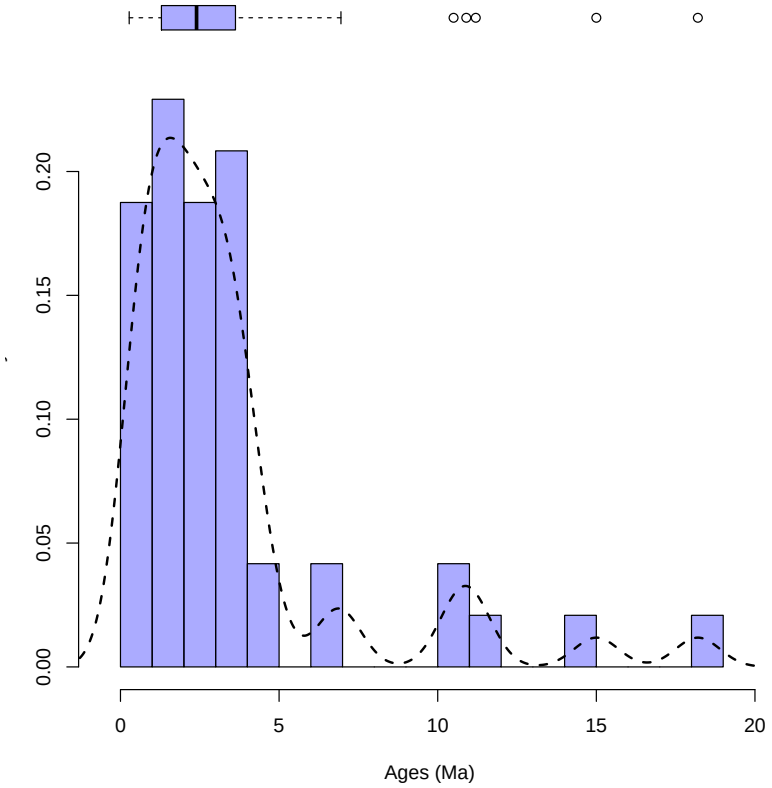

Figure 1: Descriptive statistics of the divergence time data. The histogram illustrates that the multimodality in the density line is an artifact of discontinuity in the age data, while the boxplot on top of the figure shows the existence of multiple outliers to the right of the distribution. All three methods agree in showing the asymmetry in the distribution with a heavy right tail. Discontinuous line represents the empirical probability density.

4.1 Concatenation and total-evidence analysis

Several tools are available for concatenating matrices, although few are able to deal with concatenating different data types. The function `concatNexus` was designed for including both types of matrices, DNA and morphological (also called standard in nexus format). Suppose we have several DNA alignments as well as a morphological partition in a directory called data, for which we want to generate a concatenated nexus file with definite partitions. Suppose now that we have our morphological matrix in some form of table, probably a spreadsheet. We can use the function `table2nexus` in order to read a delimited table file from the disk and the write back the nexus-formatted matrix:

```
# convert from comma-delimited to nexus
R> table2nexus(path="data/morpho.csv",
               datatype="standard",
               header=FALSE,
               sep=",",
               con="data/morpho.trimmed.nex")
```

Now that the morphological matrix is in nexus format, we can concatenate the DNA and morphology alignments follows:

```
library(tbea)

# our alignments are in a directory called data
R> aligns <- dir(path="data", pattern=".nex")
R> aligns <- paste("data/", aligns, sep="")

# concatenate alignments indicating which of them is
# a morphological matrix
# the function will write to dist the file in filename and
# a file with the partition information in partitions.txt
R> concatNexus(matrices = aligns,
               filename = "concatenatedMolmorph.nexus",
               morpho = TRUE,
               morphoFilename = "data/morpho.trimmed.nex",
               sumFilename = "partitions.txt")
# also, we can provide a pattern and a path and the function
# will find the alignments and concatenate itself
R> path <- "data"
R> pattern <- "trimmed.nex$"

R> concatNexus(pattern = pattern,
               filename = "concatenatedMolmorph.nexus",
               path = path,
               morpho = TRUE,
               morphoFilename = paste(path, grep(pattern = "morpho.",
                                                 x = dir(path, pattern),
                                                 value = TRUE),
                                      sep = "/"),
               sumFilename = "partitions.txt")
```

This will generate two files, one with the concatenated nexus alignment (`concatenatedMolMorph.nexus`) and one with information on the start and end position for each partition (`partitions.txt`).

4.2 Calibration specification

Two calibration priors can be implemented in the divergence time analysis of the family Cynodontidae, one using node calibration and one with tip calibration.

4.2.1 The genus *Hydrolycus*

Geological context The oldest record of this genus has been found in the Pozo formation of Peru (Antoine et al., 2016). The locality CTA-51 is located in the Contamana region, on the north-eastern flank of the Maquia anticline. CTA-51 has been reported as "late middle Eocene (early Barrancan, >41.6-40.94 Ma)" in the supplementary material of Antoine et al. (2016).

Calibration position Although most of the fossil record of this family is best identified as cynodontid indeterminate specimens, rather than to represent the genus *Hydrolycus* as suggested by Antoine et al. (2016). The implication of this decision impacts divergence time analysis since while considering it to represent *Hydrolycus* puts higher probability on the origin of such genus to be closer to the age of the Contamana specimens. On the other hand, to use them to calibrate the node of the family would put higher probability of the whole family originating closer to the time of these specimens. It is unlikely that the fossil remains of Contamana are closer in geologic time to the origination time of the family in comparison to being closer to the origination in geologic time for the genus *Hydrolycus*. If used to calibrate the node of the family (i.e., the MRCA of *Rhaphiodon vulpinus* and *Hydrolycus armatus*), the estimated posterior ages would be too young and close to the node calibration, while if it is used to calibrate the node of the genus *Hydrolycus* (i.e., the MRCA of *Hydrolycus armatus* and *H. scomberoides*) this bias in time estimation would be weaker.

Node calibration density Given the age uncertainty mostly based on position and biostratigraphy and that no specific radiometric estimate applies closer in the stratigraphic context to this locality, it is best to use a uniform distribution with first and last time parameters 40.94 and 41.6 Ma respectively $t_{Hydrolycus} \sim U(40.94, 41.6)$. There is however an alternative which is to consider a distribution with soft bounds. A popular alternative is the lognormal distribution, and we can find the combination of mean and standard deviation that better describes a distribution whose density values 0.0, 0.5, and 0.975 correspond to the quantiles defined by the uniform prior above. As the standard lognormal distribution is defined between $(0, \infty)$, we need to apply an offset towards the minimum age, 40.94 in this case. This will make it possible to relax the rather strong assumption about the node age, because by using a uniform prior the probability of observing times with fall outside the interval is zero. In contrast, the lognormal distribution will allow for soft bounds around the calibration density. The function `findParams` can be used to find such combination of parameters for a given pair of quantiles as follows:

```
# we need to substract the minimum to each quantile because
# we want to offset later
R> findParams(q = c(40.94-40.94, 41.27-40.94, 41.60-40.94),
              p = c(0.0,  0.50, 0.975),
              output = "complete",
              pdfunction = "plnorm",
              params = c("meanlog", "sdlog"), initVals=c(1,1))
$par
[1] -1.1085098  0.3538003

$value
[1] 3.501584e-08

$counts
```

```
function gradient
      89       NA

$convergence
[1] 0

$message
NULL
```

The function recovers the mean (in log scale) as -1.1085098 and the standard deviation as 0.3538003. We can therefore define in `Beast2` a calibration density $LN(-1.1085098, 0.3538003)$ (setting `meanInRealSpace` to false in `beauti`) in the divergence time estimation analysis. We can further plot the lognormal density as it would be used by Beast2 with the function `lognormalBeast` (Figure 2; please note that the values from and to are counted in units from the offset, not in real scale):

```
R> plot(lognormalBeast(-1.1085098, 0.3538003, meanInRealSpace=FALSE,
                       offset=40.94, from=0, to=4),
        type="l", lwd=2, main="Node calibration for Hydrolycus",
        xlab="Time (Ma)")
```

4.2.2 The specimen *Hydrolycus* MUN 16211 as tip calibration

The fossilized birth-death process as expanded through sampled ancestors (Heath et al., 2014; Gavryushkina et al., 2014) uses a uniform distribution on the tip ages for both extant and extinct taxa. Thus, for all data from extant taxa a value of 0 Ma is going to be used while for the fossil specimen *Hydrolycus* MUN 16211 an interval of 2.58–5.33 corresponding to the age interval of the Pliocene will be used as model for this tip. Further configurations are unnecessary as this taxon has morphological data that make unnecessary to constrain its position in the tree.

```
### exponential prior with mean in Ware
R> findParams(q = c(0, 3.96),
              p = c(0.0,  0.50),
              output = "complete",
              pdfunction = "pexp",
              params = c("rate"), initVals=c(1))
# plot the calibration density
curve(dexp(x, rate=0.17), from =0, to=10,
      main = "Tip calibration for\nthe fossil",
      xlab = "Time (Ma)", ylab="Density")
```

It is important to note that the exponential distribution can be parametrized in terms of the mean ($1/\lambda$, as is the case for `Beast2`) or the rate (λ) as used by R. Therefore the value to input when defining the prior of an exponential distribution in `beauti` is, which in our case is $1/0.17 = 5.8$. Note however, that when plotting the prior in R (e.g. using the function `curve` as above) we need to use λ (i.e., 0.17) rather than the mean (Figure 3).

4.3 Initial trees

In a divergence times estimation, initial trees are used in three different ways: (i) when the topology and divergence times are co-estimate, the MCMC sampling starts by proposing an initial random tree (e.g., `Beast2`); (ii) when complex analyses are carried out comprising several terminals and multiple calibration points, it is possible to provide a previously estimated initial tree which makes it easier

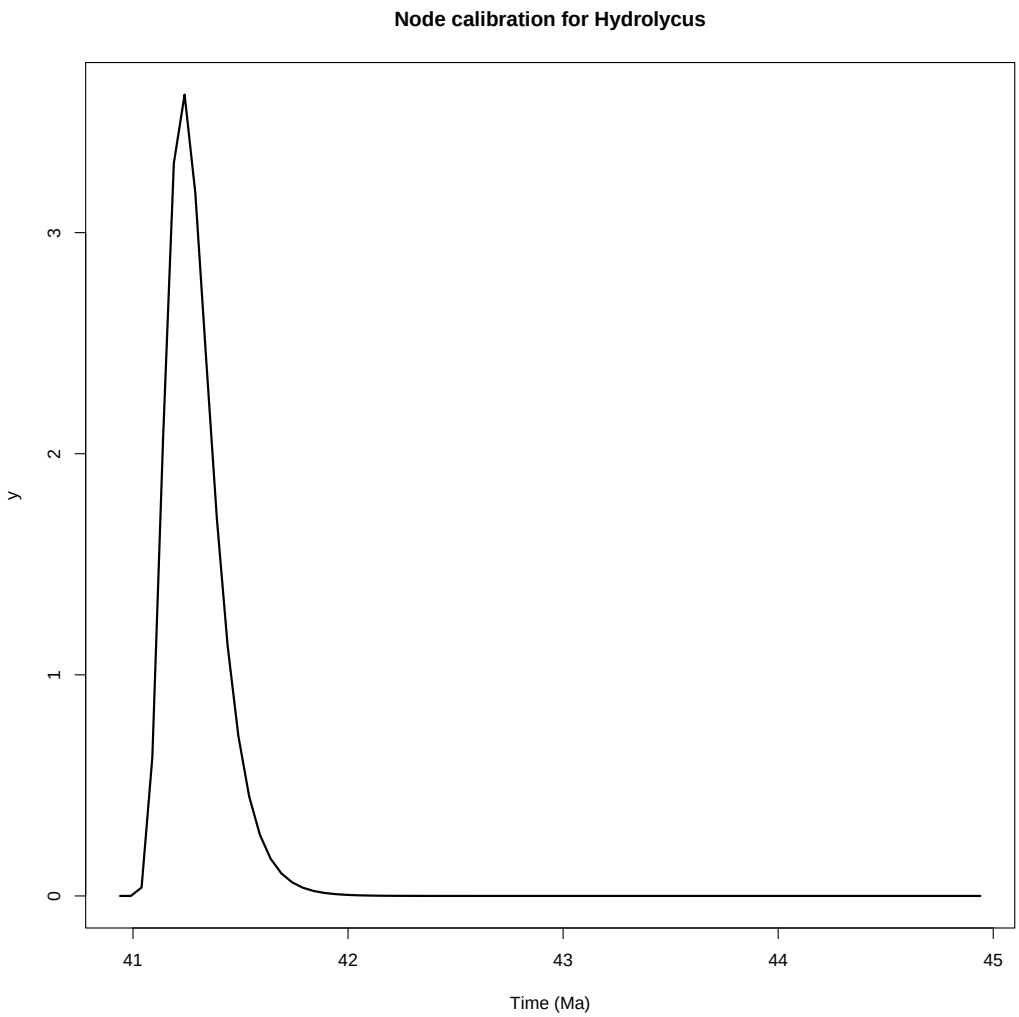

Figure 2: Calibration density for the node corresponding to the genus *Hydrolycus* as generated by the function `lognormalBeast`.

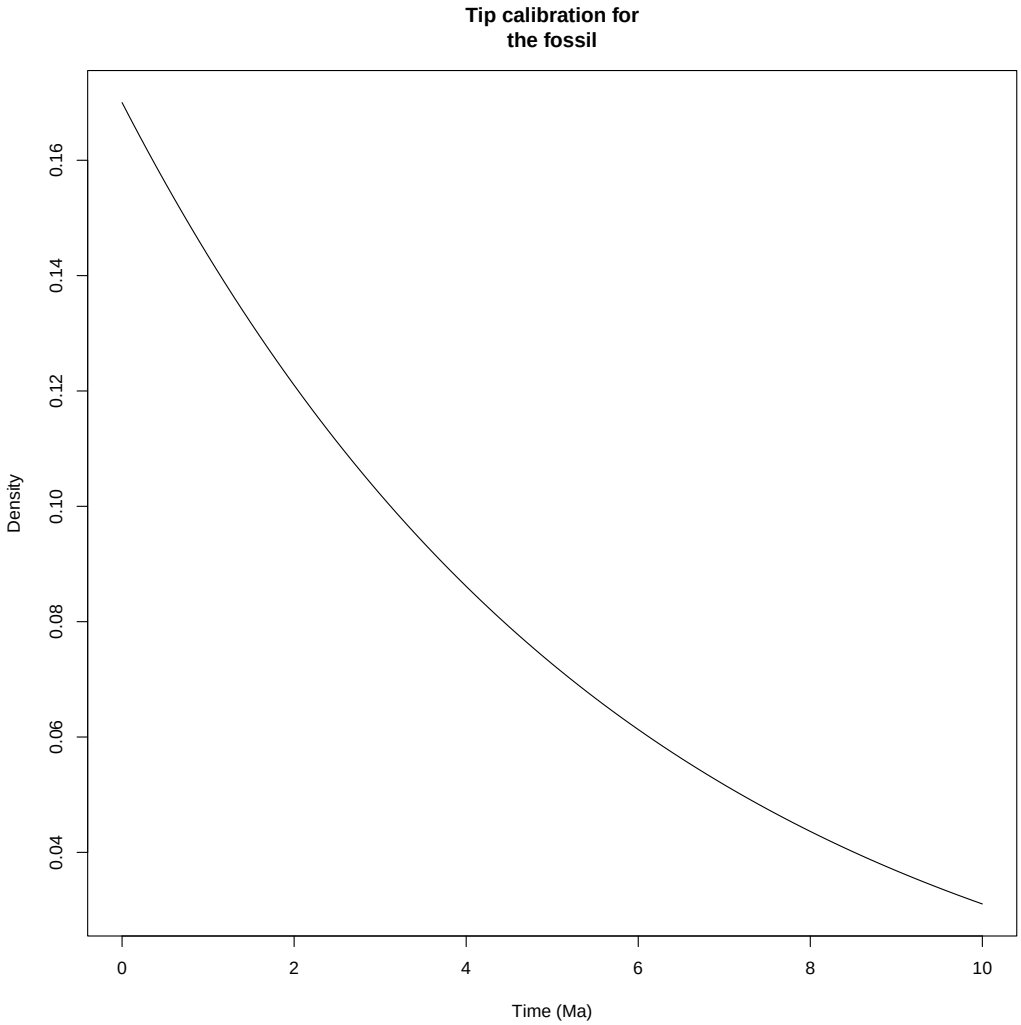

Figure 3: Calibration density for the node corresponding to the genus *Hydrolycus* as generated by the function `lognormalBeast`.

to reach convergence; and (iii) when we want to fix the topology and just estimate divergence times. However, node times must be compatible with the calibration densities, otherwise, the likelihood of the initial tree is zero, and thus, their posterior probability. This generates an issue in programs such as `Beast2` which will try a finite (say 100) number of tree proposals; if all these fail, the analysis will not start.

Initial trees are challenging to prepare by hand because they need to be consistent with all node calibrations so that the product of likelihood and priors is different from zero. Also, the tree needs to be in newick format, and the larger the tree, the more complicated the newick string. Node times need to be within the domain of the calibration distribution, so that branch length need to be specified accordingly.

We can produce initial trees by hand, or infer them using a fast method such as maximum likelihood or maximum parsimony. If the latter is used for inferring an initial tree, `TNT` (Goloboff et al., 2008) is a popular alternative in use. However, its format is specific to the program and differs from standard newick, so that in can be impossible to use it as initial tree in other types of analyses. `tbea` has a function for converting from TNT tree format to standard newick, thus allowing to use parsimony trees as initial proposals in Bayesian programs.

```
# create a file with multiple trees tree in TNT format to convert
# to newick format
R> writeLines(
        text = c(
        "tread 'three arbitrary trees in TNT format'",
        "(Taxon_A ((Taxon_B Taxon_C)(Taxon_D Taxon_E)))*",
        "(Taxon_A (Taxon_B (Taxon_C (Taxon_D Taxon_E))))*",
        "(Taxon_A (Taxon_C (Taxon_B (Taxon_D Taxon_E))));",
        "proc-;"
        ),
        con = "someTrees.tre"
    )

R> tbea::tnt2newick(file = "someTrees.tre", return = TRUE)
[1] "(Taxon_A,((Taxon_B,Taxon_C),(Taxon_D,Taxon_E)));"
[2] "(Taxon_A,(Taxon_B,(Taxon_C,(Taxon_D,Taxon_E))));"
[3] "(Taxon_A,(Taxon_C,(Taxon_B,(Taxon_D,Taxon_E))));"

R> file.remove("someTrees.tre")
```

In order to generate a tree in which node times are consistent with the calibration priors we can use a fast method such as penalized likelihood (implemented in the `ape` package) using the initial tree and our set of node calibrations:

```
R> tree <- ape::read.tree(text="(Acestrorhynchus_falcirostris:0.1296555,
     ((((Cynodon_gibbus:0.002334399,Cynodon_septenarius:0.00314063)
     :0.07580842,((Hydrolycus_scomberoides:0.007028806,Hydrolycus_MUN_16211
     :0.009577958):0.0225477,(Hydrolycus_armatus:0.002431794,
     Hydrolycus_tatauaia:0.002788708):0.02830852):0.05110053)
     :0.02656745,Hydrolycus_wallacei:0.01814363):0.1712442,
     Rhaphiodon_vulpinus:0.04557667):0.1936416);")
## get the node IDs according to the mrca of pairs of appropriate species
# root
R> mrca(tree)["Acestrorhynchus_falcirostris", "Hydrolycus_armatus"]
[1] 10
# Hydrolycus sensu stricto (cf. Hydrolycus)
```

```
R> mrca(tree)["Hydrolycus_scomberoides", "Hydrolycus_armatus"]
[1] 15
# Hydrolycus scomberoides (including the fossil)
R> mrca(tree)["Hydrolycus_scomberoides", "Hydrolycus_MUN_16211"]
[1] 16
# specify the calibrations as a data frame
R> calibs <- data.frame(node = c(10, 15, 16),
                        age.min = c(43, 40.94, 0),
                        age.max = c(60, 41.6, 5.33),
                        stringsAsFactors = FALSE)
# calibrate the tree with the information in calibs
R> set.seed(15)
R> calibrated <- ape::chronos(phy = tree, calibration = calibs)

# plot the calibrated tree
R> plot(calibrated, label.offset = 1.2)
R> axisPhylo()

# fetch the plotted tree from the graphic environment
R> plttree <- get("last_plot.phylo", envir=.PlotPhyloEnv)

# process the node heights and reorder as needed
R> nodes <- cbind(plttree$xx, plttree$yy)
R> nodes <- nodes[c(10, 15, 16),]
R> nodes <- cbind(nodes, calibs[, c("age.min", "age.max")])
R> colnames(nodes) <- c("x", "y", "min", "max")
R> nodes <- cbind(nodes, rev.min = -(nodes[, c("min")] - max(plttree$xx)))
R> nodes <- cbind(nodes, rev.max = -(nodes[, c("max")] - max(plttree$xx)))

# plot the original calibrations as error bars
R> arrows(x0=nodes[, "rev.max"], x1=nodes[, "rev.min"],
          y0=nodes[,"y"], y1=nodes[,"y"], angle=90, length=0.3,
          lwd=3, code=3, col="darkblue")

R> nodelabels()
R> tiplabels()

# write the tree
R> write.tree(phy = calibrated, file = "startingTree.newick")
```

4.4 Post-analysis comparisons

4.4.1 Summary of (posterior) tree distributions

There are two steps necessary for summarizing tree samples. One is to decide how many distinct topologies (disregarding branch lengths) are in the collection of trees, and how many trees fall within each category. Then we need to summarize branch lengths for each of these tree subsets using any sensible summary measure (e.g., mean, median). The functions `topoFreq` and `summaryBrlen` aid in these two tasks respectively. One important point when combining these two functions is that we need to decide carefully whether the trees are rooted or unrooted. The reason for this is that `topoFreq` uses the Robinson-Foulds (RF) distance among trees in order to find clusters of trees which are identical in topology and then bind them together in a list so that we can operate over them later. As the RF

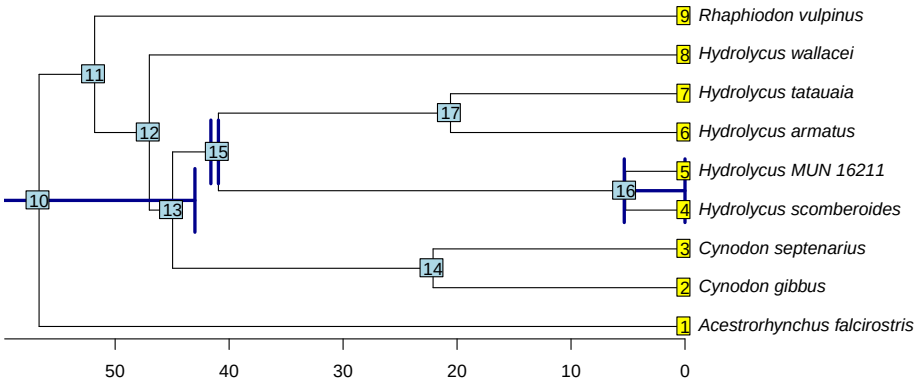

Figure 4: Calibrated tree which is consistent with the calibration densities and can be used as starting tree in analyses of divergence time estimation.

distance is defined over unrooted trees, we would need to unroot them if they are originally rooted in order to use the function. However, unrooting will destroy the root node and then add up together both branch lengths incident onto it, and this will be reflected when summarizing with `summaryBrlens` using the output of `topoFreq` as input. This will return one branch length less than the number present in the original sample of rooted trees. Accordingly, this method will only work when collections of unrooted trees, or collections of time trees of fixed topology where we are only interested in finding a summary values for branch lengths. It should not be used when both the topology and divergence times are being coestimated (e.g., when the posterior of trees comes from not fixing the topology in `Beast2`, `RevBayes`, and `MrBayes`) but it is perfectly fine when the inference was done to estimate only the topology and branch lengths, such as in standard tree inference in `MrBayes`. The reason is that the inferential process in the latter works with unrooted trees. We see an example using these two functions below where we want to use the median branch lengths on the topology for each group:

```
R> library(ape)
R> library(tbea)

# get the data from https://github.com/gaballench/cynodontidaeWare
R> trees1 <- "cynodontidaeWare/bayesian/mrbayes/concatenatedMolmorph.nexus.run1.t"
R> trees2 <- "cynodontidaeWare/bayesian/mrbayes/concatenatedMolmorph.nexus.run2.t"
R> mtr1 <- read.nexus(trees1)
R> mtr2 <- read.nexus(trees2)

# add together the two tree samples
R> trees <- c(mtr1, mtr2)

# calculate the topological frequencies
R> tpf <- topoFreq(trees, output="trees")

# summarise brach lengths using the median
R> sumtrees <- summaryBrlen(tpf$trees, method="median")
```

```
# create an index vector to
# sort the frequencies in decreasing order
R> decreasingIdx <- order(tpf$fabs, decreasing=TRUE)

# how many topologies comprise about 90% of the trees in the sample?
R> cumsum(tpf$fabs[decreasingIdx])/sum(tpf$fabs)
[1] 4

# plot these four trees with branch lengths already summarised
R> par(mfrow=c(2,2))
R> plot(sumtrees[[decreasingIdx[1]]], type="unrooted", show.node.label=FALSE,
        cex=1.5, main=round(tpf$frel[decreasingIdx[1]], digits=2))
R> plot(sumtrees[[decreasingIdx[2]]], type="unrooted", show.node.label=FALSE,
        cex=1.5, main=round(tpf$frel[decreasingIdx[2]], digits=2))
R> plot(sumtrees[[decreasingIdx[3]]], type="unrooted", show.node.label=FALSE,
        cex=1.5, main=round(tpf$frel[decreasingIdx[3]], digits=2))
R> plot(sumtrees[[decreasingIdx[4]]], type="unrooted", show.node.label=FALSE,
        cex=1.5, main=round(tpf$frel[decreasingIdx[4]], digits=2))
```

4.4.2 Prior-posterior comparisons

After carrying out a divergence time analysis, in this example with `Beast2`, we might be interested in comparing the prior and the posterior of node ages. We can use the function `crossplot` to generate a plot where the x-axis represents one analysis and the y-axis another, for instance, prior versus posterior:

```
R> prior_path <- "prior/cynoDivtime1.log"
R> posterior_path <- "posterior/cynoDivtime1.log"
R> crossplot(log1Path=prior_path,
             log2Path=posterior_path,
             skip.char="#",
             pattern="mrca.date",
             bar.lty=1,
             bar.lwd=2,
             identity.lty=2,
             identity.lwd=2,
             main="Prior-posterior comparison\nCynodontidae",
             xlab="Prior node age (Ma)",
             ylab="Posterior node age (Ma)",
             pch=20, cex=2, col="blue")
```

This kind of plot has been used in the literature when comparing prior and posterior MCMC samples, as well as when comparing the same kind of estimates coming from different independent runs or types of analysis (e.g., Drummond and Stadler, 2016; Álvarez-Carretero et al., 2022). The plot generated (Figure 6) allows to see clearly that the prior node densities have more variance than the posterior ones, which is a consequence of using the data through the likelihood function.

The function `measureSimil` integrates the area under the curve defined as the intersection between both distributions. It is a descriptive measure of how similar two distributions are. The function can both plot the resulting distributions and their intersection (Figure 7) as well as print out its value, or skip the plot and just return the value:

```
R> priorData <- read.delim(prior_path, comment.char="#")
R> posteriorData <- read.delim(posterior_path, comment.char="#")
```

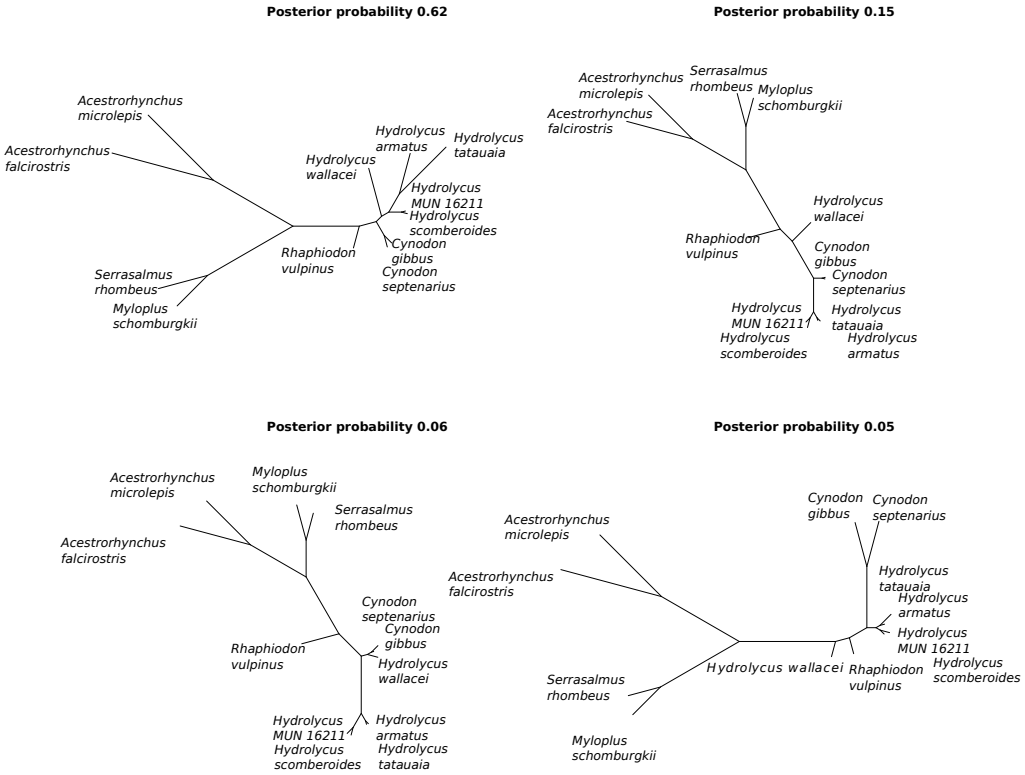

Figure 5: Summary of a sample of trees which comprise nearly 90% of the tree posterior sample. These are labeled by posterior probability and sorted in decreasing order. Note the dramatic difference in branch lengths between topologies represented in a small frequency when compared with the topology which was present most frequently: The branch lengths estimates in these small samples are unreliable whereas the ones in larger samples are more reliable.

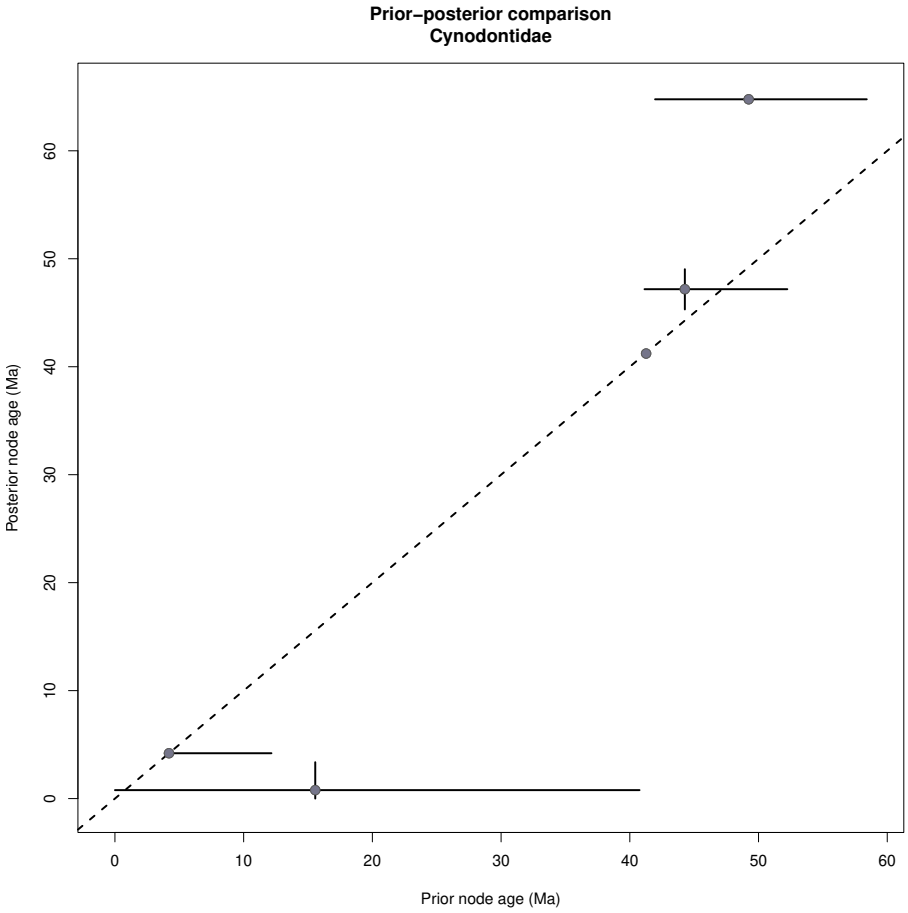

Figure 6: Comparison between the prior and the posterior of node ages in the divergence time estimation of the Cynodontidae dataset.

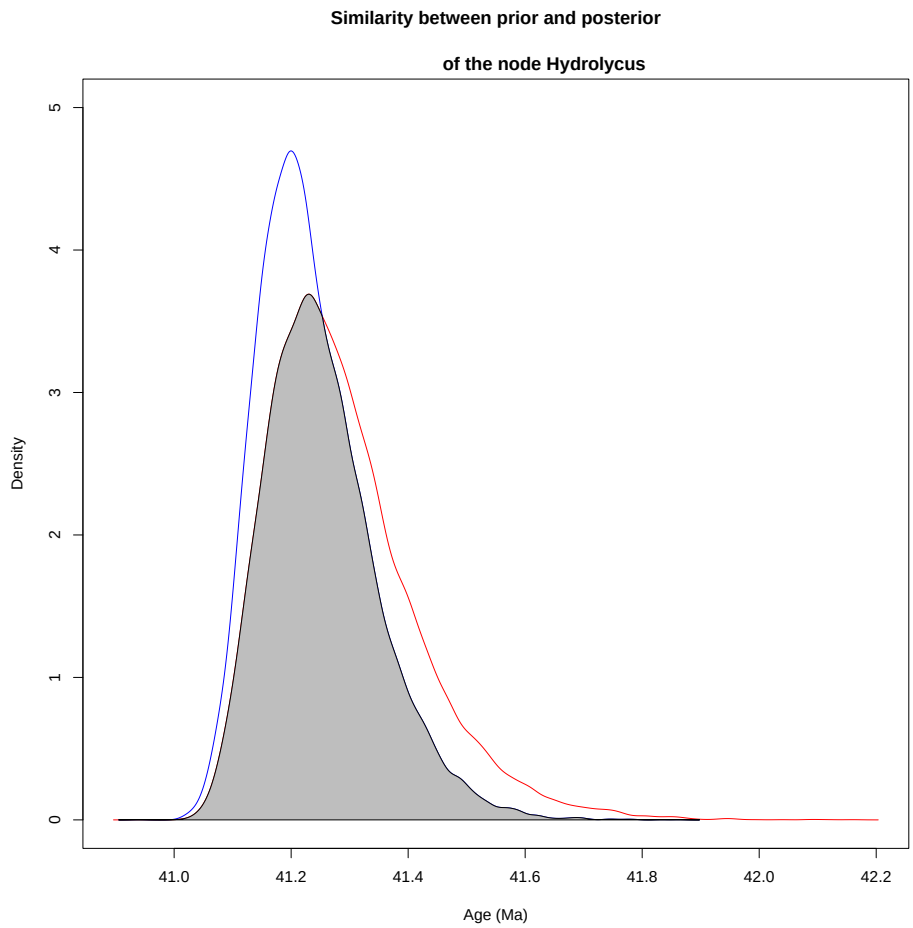

Figure 7: Similarity between the prior (red) and posterior (blue) densities of the node *Hydrolycus*. The value of the intersection between both distributions is 0.83.

```
R> measureSimil(d1=priorData$mrca.date.backward.Hydrolycus.,
                d2=posteriorData$mrca.date.backward.Hydrolycus.,
                ylim=c(0, 5),
                xlab="Age (Ma)",
                ylab="Density",
                main="Similarity between prior and posterior\n
                      of the node Hydrolycus")
[1] 0.8288267
```

4.5 Inference of origination time given a collection of time estimates

Different estimates of the same quantity are often of interest in meta-analyses as well as means of specifying secondary node calibrations. However, how to summarize such information is not straightforward (Hill, 2011). Here we implement three ways in which estimates from different distributions measuring the same quantity can be summarized in a single distribution: (i) using stratigraphic intervals, (ii) x-intercept confidence intervals, bootstraping the distribution of mean and median, and (iii) conflation of distributions. The package `tbea` has implementations of each of these methods.

4.5.1 Stratigraphic intervals

Some methods are available for estimating the confidence intervals on the extreme points of a stratigraphic interval, i.e. an estimation under a certain level of confidence for the point of appearance and disappearance of a given taxon in geologic time. These methods are relevant here since we can think of a biogeographic events in an analog way with a taxon in geologic time where each biogeographic event has a birth and death in time in the same way as taxa appear and become extinct. This approach is so far inexistent in historical biogeography and seems to fit the main goal of determining the final point of connection between biogeographic areas. Marshall (2010) summarized different approaches for estimating confidence intervals under certain sampling regimes, assumptions and confidence levels, ranging from classical to Bayesian-based estimators. The package `tbea` implements two of these methods, the classical Strauss and Sadler (1989) and the distribution-free estimator of Marshall (1994), as described below.

The first method provides estimators for one- (θ_1) or two-parameter (θ_1, θ_2) cases are constructed adding and subtracting a portion α of the magnitude of the observed stratigraphic interval $[y, z]$. Such constant depends on the confidence level wanted (e.g., 95%) and its calculation is different for each parameter case.

$$y - \alpha(z - y) < \theta_1 < y \tag{2}$$

$$z < \theta_2 < z + \alpha(z - y) \tag{3}$$

The constant α can be calculated either exactly (one-parameter case) or iteratively (two-parameter case) since the latter can not be solved for alpha analytically.

The method of distribution-free confidence intervals was developed without assuming any particular distribution of underlying gap sizes, instead, it calculates quantiles of gap size between occurrence points for an ordered vector of gap sizes. The major cost is that such intervals are larger that classical CIs and on the other hand, require much more data for constructing intervals of high confidence (e.g., 95%).

The method works constructing a confidence interval that has a level of confidence C (i.e., Cth percentile) for a γ confidence probability with the information from N gaps.

The lower bound is then the i_{th} gap in the ordered vector applied to $min(vector) - x_i$ where x_i is the magnitude of such i_{th} element in the vector. The formula is as follows:

$$\gamma > \sum_{x=0}^{x} \binom{N}{x} C^x (1 - C)^{N-x} \quad \text{for } x \leq N \tag{4}$$

The upper limit can be constructed with the formula:

$$(1 - \gamma) < \sum_{x=0}^{x} \binom{N}{x} C^x (1 - C)^{N-x} \quad \text{for } x \leq N \tag{5}$$

The method makes use of the binomial distribution in order to find the quantiles that satisfy the condition of being smaller that γ and from these picking the largest (in the case of the lower bound) and the smallest (for the upper bound). It is noteworthy that if only one x satisfies Equations (4) or (5), no lower or upper confidence interval can be constructed. Also, if $x = N$ in (5), no upper confidence interval can be constructed since there is no $(x + 1)_{th}$ gap in the vector. Marshall (1994) mentioned that inference using his method is only valid as long as there is no trend in gap size (i.e., a trend of increase or decrease in gap size as we go along the time data). Then, mentioned the transposition test (Kreyszig, 1979), as a way to determine whether there is any increasing or decreasing trend in data. There are other statistical tests which could however be applied for testing monotonic trends (e.g., the Mann-Kendall test for monotonic trend; Mann, 1945; Kendall, 1948). It is important to note that because the method works with quantiles of the sorted gap sizes, these are first calculated from

the sorted vector of times, but then internally sorted when using the method "Marshall94"; however, whether there is any monotonic trend must be assessed on the unsorted version. One possibility is to use the Mann-Kendall test on the times vector prior to applying the function:

```
R> library(Kendall) # for using the MannKendall function
R> data(andes)
R> andes <- andes$ages
# remove missing data
R> andes <- andes[complete.cases(andes)]
# remove outliers
R> andes <- sort(andes[which(andes < 10)])
R> gapsizes <- sapply(2:length(andes), FUN=function(x) andes[x] - andes[x-1])
R> Kendall::MannKendall(gapsizes)
tau = 0.119, 2-sided pvalue =0.27337
```

According to the Mann-Kendall test, we fail to reject the null hypothesis of no monotonic trend in the gap sizes. Therefore we may apply the "Marshall94" method for bracketing the confidence interval.

These two methods are available in the function `stratCI` which allows to pick one of two methods: "straussSadler89" or "marshall94". We can use the same data from the previous block in order to illustrate the usage of the function stratCI:

```
# calculate the older theta that is an extension beyond
# the oldest (first from past to present) occurrence
# using the method Strauss-Sadler89
R> stratCI(andes, method="Strauss-Sadler89", nparams="one.par", C=0.95, endpoint="first")
     minObs     maxObs     maxEst          H     alphaR      alpha
  0.2749000  6.9500000  7.4435055 43.0000000  0.4935055  0.0739323
# calculate the lower and upper bounds on the confidence interval with a quantile 0.8
R > stratCI(andes, method="Marshall94", confidence = 0.95, quantile = 0.8)
[1] 7.0701 7.2147
R > stratCI(andes, method="Marshall94", confidence = 0.95, quantile = 0.95)
[1] 7.14   NA
Warning message:
In upperBound(confidence, length(gapsizes), quantile) :
  Impossible to calculate upper bound
```

As Marshall (1994) already noted, the method requires quite a large number of gaps in order to calculate both bounds for the 95% confidence interval; for our dataset with just 42 gaps (i.e., 43 times), we can only find both bounds for it at the 80% confidence level (or lower), and thus claim that the lower and upper bounds for the 95% confidence interval are 7.07–7.2Ma. When we raise the confidence level to 95%, we can only infer the lower bound, thus being able just to claim that the 95% confidence interval is at least as old as 7.14 Ma. It is important to realize that the higher the confidence level for finding the bounds on the 95% confidence interval, the better, in a similar way as calculating a 95% confidence interval is better than calculating a 70% confidence interval.

We can plot these extensions after the oldest occurrence and compare the results of both methods (Figure 8):

```
# calculate the confidence interval on the oldest time using the
# Strauss-Sadler89 method
R> straussSadlerAndes <- stratCI(times = andes, method="Strauss-Sadler89",
                                 nparams = "one.par", endpoint = "first",
                                 C = 0.95)
#calculate the confidence interval under th distribution-free approach
```

```
R> marshallAndes <- stratCI(andes, method="Marshall94", confidence = 0.95,
                            quantile = 0.8)
# plot the estimations
R> hist(andes, probability = TRUE,
        col = rgb(red = 0, green = 0, blue = 1, alpha = 0.3),
        xlim = c(0, 10),  main = "CIs based on stratigraphic intervals",
        xlab = "Age (Ma)")
# plot the lines  for the classical CI estimator of Strauss and Sadler
R> arrows(x0 = straussSadlerAndes["maxObs"], y0 = 0.025,
          x1 = straussSadlerAndes["maxEst"], y1 = 0.025, code = 3, angle = 90,
          length = 0.1, lwd = 3, col = "darkblue")
# plot the lines for the CI estimator of Marshall 94
R> arrows(x0 = marshallAndes[1], y0 = 0.05, x1 = marshallAndes[2], y1 = 0.05,
          code = 3, angle = 90, length = 0.1, lwd = 3, col = "darkgreen")
# plot a legend
R> legend(x = "topright", legend = c("Strauss and Sadler \'89", "Marshall \'94"),
          col = c("darkblue", "darkgreen"),  lty = 1, lwd = 3)
```

4.5.2 x-intercept inference

These approaches are based on the assumption that the cumulative the pattern of appearance of divergence time events from an unknown time t_0 onwards can inform on the value of such event. Thus, a regression model with an additional parameter x-intercept may be thought of as an estimator for such point in time given the trend present in the linear model (Figure 9). Moreover, the estimation of uncertainty regions such as confidence intervals on the x-intercept can be even more informative than the point estimate itself; such regions can be constructed by fitting of empirical cumulative density function.

Events in time start at some (often unobserved) point and the their realizations continue appearing until the process finishes. Because they are delayed by some quantity after the generator process, we can use the pattern of occurrence in time in order to estimate the origination time. They can be described as a distribution regardless of its generator process, which will follow a probability density function (PDF) $f(t)$, with an associated cumulative density function (CDF) $P(X < x)$. We can fit an empirical CDF for which there are several methods available given the properties of CDFs. One option of curve fitting is the linear regression. The only case where a CDF would be truly linear would be in the case of a uniform distribution where

$$P(X \leq x) = \frac{x}{b-a} \text{ for } x \in [a, b]$$

which implies that the empirical data in time follow a uniform distribution. For any other case, a fitted CDF would be a non-linear, and thus a linearization should be applied to it.

There are several methods available for linearizing data (e.g., applying the logarithm, the square and cubic roots, the arcsine, or the reciprocal to the raw data). In the case of CDF fitting, we need to guarantee that the data are linear regardless of the PDF in order to estimate parameters of the model and thus the procedure is appropriate. After model fitting, a back transformation would be necessary for representing the time value of interest in real space.

The linear model has the form

$$y = mx + b$$

where m is the slope and b is the y-intercept. However, we are not interested in any of these parameters but instead in a third one, the x-intercept, that is defined as the point in x where $y = 0$. This parameter can be defined by letting $y = 0$ in the equation and thus

$$x_{y=0} = -\frac{b}{m}$$

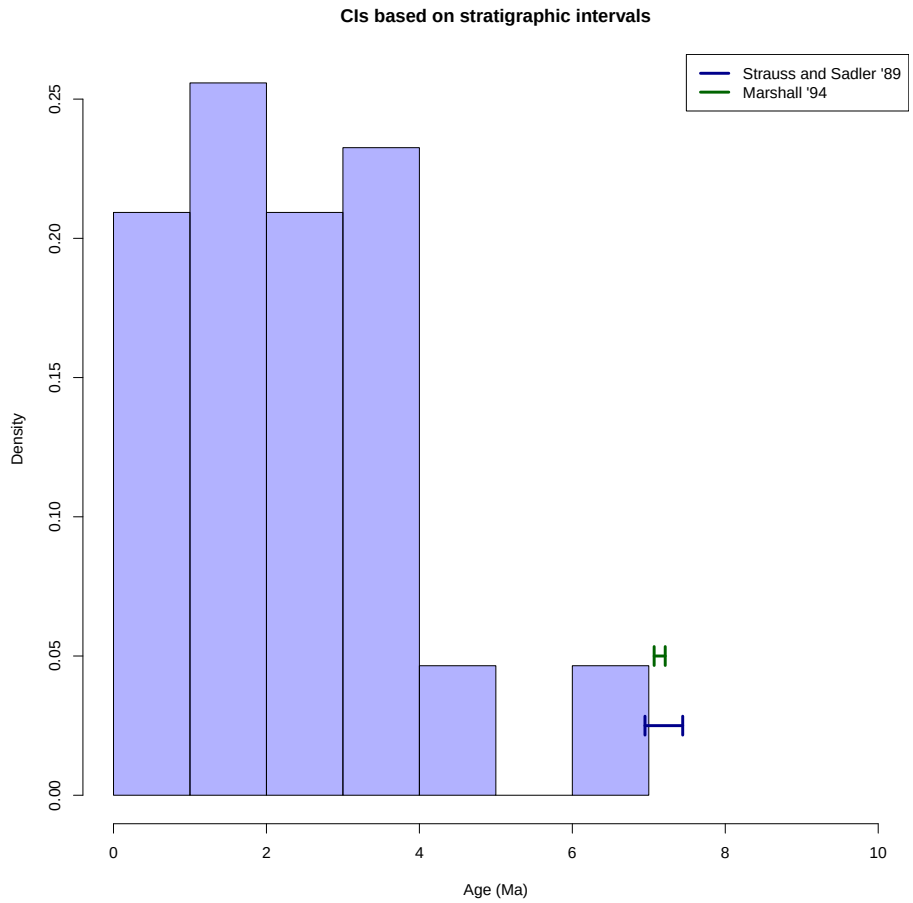

Figure 8: Estimation of the time of separation between cis- and trans-Andean drainages based on the methods of stratigraphic intervals of Strauss and Sadler (1989) and Marshall (1994).

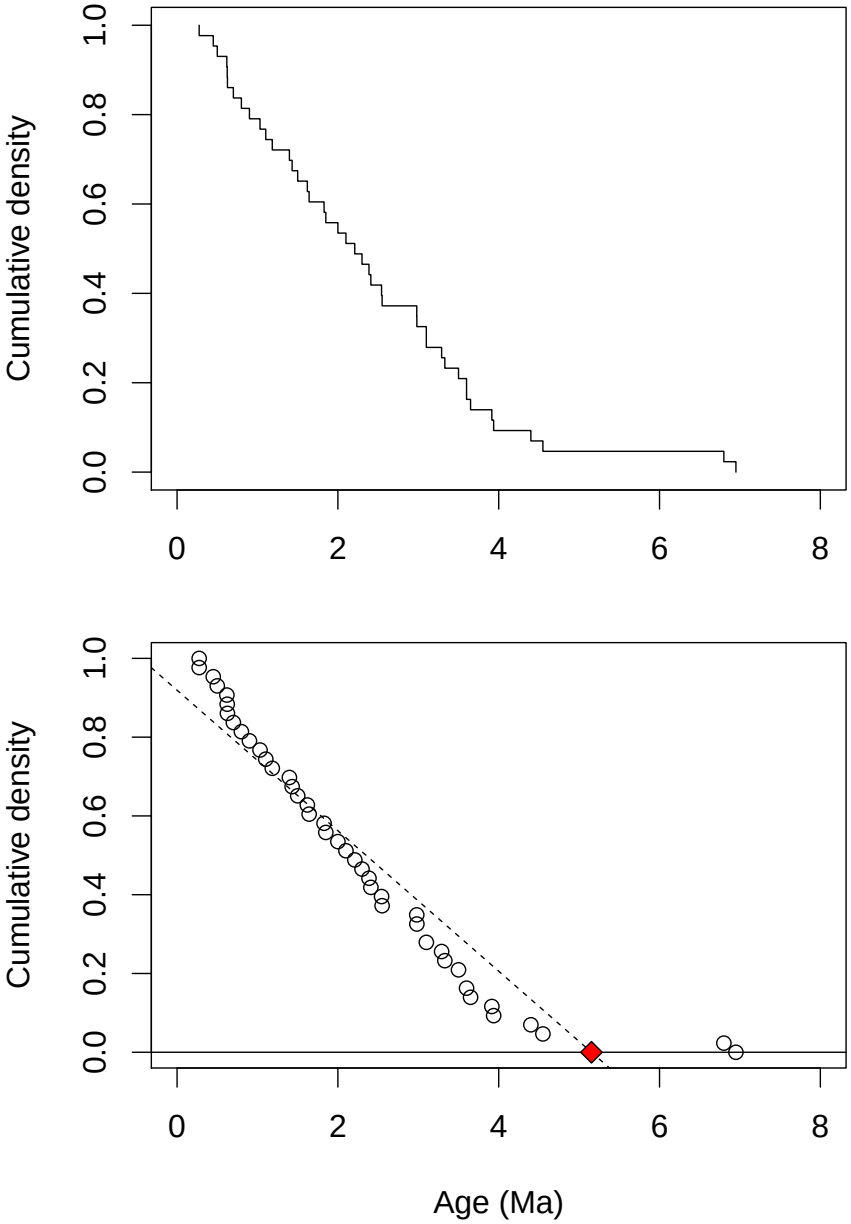

Figure 9: Empirical cumulative distribution of divergence time events. Above, plot of the empirical density; below, model fit of the cumulative distribution. The interrupted line represents the regression model on the cumulative values, and the red diamond represents the x-intercept.

allows us to find the point at which the data began to be generated.

Several methods have been proposed for estimating the uncertainty on the x-intercept estimate, from the intuitive inverse regression $X \sim Y$ to bootstraping methods. According to Seber and Lee (2013, pp. 146-147) this problem has long been discussed in the statistical literature without much consensus, although some specific alternatives are of interest due to their theoretical appeal, ease of implementation, or simplification of assumptions. Uncertainty on the estimate of the x-intercept can be obtained through Taylor series simplification for symmetric intervals. Another approach by Draper and Smith (1998, pp.83-86) allows the calculation asymmetric confidence intervals by projecting the values of the confidence region around the linear model on the x-axis by manipulation of the curves describing the confidence region, what gives

$$CI_{X_0} = \hat{X}_0 + \frac{(X_0 - \bar{X})g \pm (ts/b_1)\{[(\hat{X}_0 - \bar{X})^2 S_{XX}] + (1-g)/n\}^{1/2}}{1-g}$$

where $g = t^2 s^2/(b_1^2/S_{XX})$, b_1 is the slope, S_{XX} is the sum of squares of X, t is the t-statistic with n-2 degrees of freedom and at $1-\alpha/2$ significance level, s^2 is the standard deviation, $\hat{X}_0$ is the estimation of the x-intercept, and $\bar{X}$ is the mean of X. Here we need to fit a linear model in order to use the curves for the confidence region, and we can use classical least square models or robust non-parametric regression (Kloke and McKean, 2014).

It is also possible to use bootstrapping in order to estimate the variance and thus the confidence interval in asymmetric cases such as the present one. Instead of using the estimates from a single fitted model on the coordinates of the empirical CDF, this method fits multiple models with random subsamples in order to construct a collection of estimated values for a parameter of interest, in this case, the x-intercept and this provide a confidence interval on the parameter.

The package `tbea` implements estimation of the confidence interval for the $x-$intercept using the method of Draper and Smith (1998) as well as estimation via bootstrap, both requiring a linear model which can be set to one of the following: using either an ordinary least squares linear model, or a robust linear model following Kloke and McKean (2014).

```
# Draper-Smith, OLS
R> draperSmithNormalX0 <- xintercept(x = andes, method = "Draper-Smith",
                                     alpha = 0.05, robust = FALSE)
# Draper-Smith, Robust fit
R> draperSmithRobustX0 <-xintercept(x = andes, method = "Draper-Smith",
                                    alpha = 0.05, robust = TRUE)
# Bootstrap, OLS
R> bootstrapNormalX0 <- xintercept(x = andes, method = "Bootstrap",
                                   p = c(0.025, 0.975), robust = FALSE)
# Bootstrap, Robust fit
R> bootstrapRobustX0 <- xintercept(x = andes, method = "Bootstrap",
                                   p = c(0.025, 0.975), robust = TRUE)

# plot the estimations
R> hist(andes, probability = TRUE,
        col = rgb(red = 0, green = 0, blue = 1, alpha = 0.3),
        xlim = c(0, 10), main = "CDF-based on confidence intervals",
        xlab = "Age (Ma)")
# plot the lines for the estimator of Draper and Smith using lm
R> arrows(x0 = draperSmithNormalX0$ci["upper"], y0 = 0.025,
          x1 = draperSmithNormalX0$ci["lower"],
          y1 = 0.025, code = 3, angle = 90, length = 0.1,
          lwd = 3, col = "darkblue")
```

```
# plot the lines for the estimator of Draper and Smith using rfit
R> arrows(x0 = draperSmithRobustX0$ci["upper"], y0 = 0.05,
          x1 = draperSmithRobustX0$ci["lower"],
          y1 = 0.05, code = 3, angle = 90, length = 0.1,
          lwd = 3, col = "darkgreen")
# plot the lines for the estimator based on bootstrap
R> arrows(x0 = bootstrapRobustX0$ci["upper"], y0 = 0.075,
          x1 = bootstrapRobustX0$ci["lower"],
          y1 = 0.075, code = 3, angle = 90, length = 0.1,
          lwd = 3, col = "darkred")
# plot a legend
R> legend(x = "topright", legend = c("Draper and Smith with lm",
                                     "Draper and Smith with rfit",
                                     "Bootstrap on x0"),
          col = c("darkblue", "darkgreen", "darkred"),
          lty = 1, lwd = 3)
```

These novel methods have produced results in line with those based on stratigraphic intervals. However, they allow estimated x-intercepts and consequently their confidence intervals that are not necessarily older than the oldest divergence time estimation event. Thus, these methods are robust to the assumption of measurement without error that is implicit in the methods based on stratigraphic intervals. Nevertheless, they are still sensitive to the presence of strong outliers as preliminary analyses showed that the regressions on these data with outliers affected strongly the robust regressions.

Both methods produce similar results (Figure 10). Although the method based on robust regression produced a CI wider than the one based on ordinary least squares, the bootstrap CI estimate was as wide as the two former combined. There is no mathematical relationship between the magnitude of such confidence intervals, but the fact that the bootstraping based on robust estimates is of more general application, it seems more adequate for these situations. It is noteworthy that these inferences are based on the assumption that the events in time are measured without error. However, this assumption relaxes due to the effect of re-sampling and uncertainty in the estimates. According to both estimators, the separation between cis- and trans-Andean areas took place between ca. 4.3–5.8 Ma.

4.5.3 Conflation of distributions

Combining different distributions into a single one is a useful tool for summarizing different sources of information. Given a collection of N distributions describing the same parameter, we aim at building a general distribution which incorporates information from the individual components. Multiple alternatives are available, and as with x-intercept estimation, the problem of summarizing a collection of distributions has several possible answers (Genest and Zidek, 1986). Hill (2008) discussed these methods and proposed the conflation as a better and general way of combining distributions describing the same parameter. Let $Q = \&(P_1, \ldots, P_N)$ be the conflation of N densities describing the same time event τ:

$$Q(\tau; \Theta) = \frac{\prod_{i=1}^{N} P(\tau_i; \Theta_i)}{\int_{-\infty}^{\infty} \prod_{i=1}^{N} P(\tau_i; \Theta_i) \mathrm{d}\tau} \tag{6}$$

Among the useful properties of conflation we have that Q is also a distribution, which can be used for calculating probabilities. Also, the conflation combines the variance of each distribution, so that intuitively densities with more uncertainty (i.e., higher variance) will have a lower impact on the conflated distribution than others with lower variance, thus weighting the informativeness of all individual distributions appropriately.

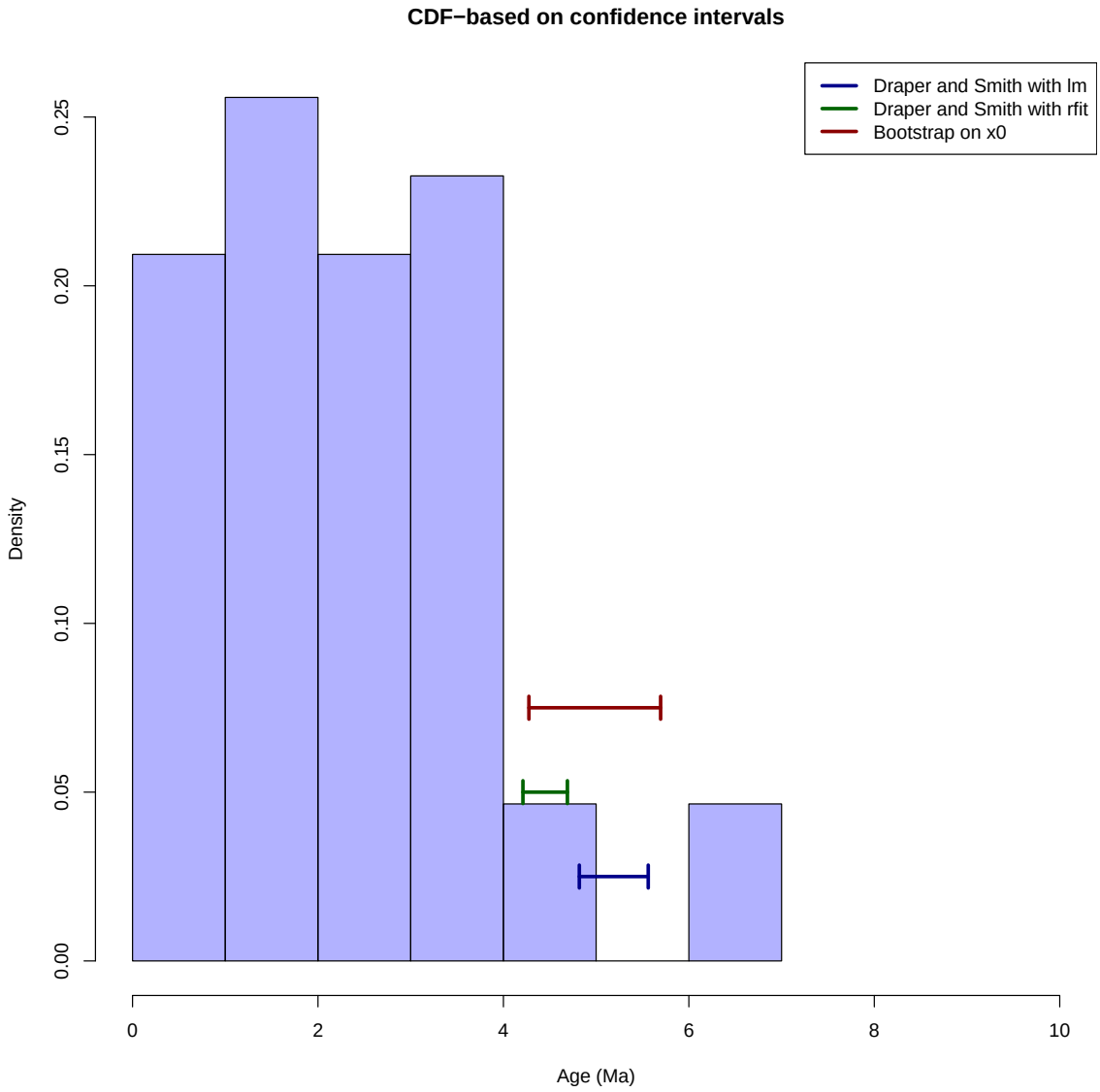

Figure 10: Estimation of the time of separation between cis- and trans-Andean drainages based on the methods of CDF fitting using the estimator of Draper and Smith (1998) using two different approaches at model fitting, one using least squares traditional regression, and the other using robust regression (Kloke and McKean, 2014) and bootstraping techniques. Both confidence intervals are near but no bound by the oldest occurrence as in stratigraphic-interval-based approaches.

Conflation proves useful when building secondary calibrations in the presence of more than two densities available for the same time event. Let us suppose that we have a biogeographic event with time τ, for which three different studies provide three independent densities, for instance, posterior probabilities for a node which corresponds to the biogeographic event. The best summary of these independent pieces of information is a conflation $Q(\tau) = \&(P_1(\tau), P_2(\tau), P_3(\tau))$, which can be used as a secondary calibration point for a new divergence time estimation analysis as it is itself a PDF. Suppose that $P_1 = N(10, 1)$, $P_2 = N(13, 0.5)$, and $P_3 = N(14.4, 1.7)$. We can calculate the conflation of these three distributions as follows:

```
# calculate the conflation and plot it
# mind the double quotes which must be used when specifying each
# distribution and the single quotes when specifying the distribution function name
R> conflate(c("density_fun(x, 'dnorm', mean=10, sd=1)",
              "density_fun(x, 'dnorm', mean=13, sd=0.5)",
              "density_fun(x, 'dnorm', mean=14.4, sd=1.7)"),
            from=0, to=20, n=101, add=FALSE, plot=TRUE)

# plot the individual distributions
R> curve(density_fun(x, 'dnorm', mean=10, sd=1), add=TRUE, col="red")
R> curve(density_fun(x, 'dnorm', mean=13, sd=0.5), add=TRUE, col="blue")
R> curve(density_fun(x, 'dnorm', mean=14.4, sd=1.7), add=TRUE, col="darkgreen")

# legend
R> legend(x="topleft",
          legend=c("Normal(10,1)",
                   "Normal(13,0.5)",
                   "Normal(14.4,1.7)",
                   "Conflation of normals"),
          col=c("red", "blue", "darkgreen", "black"),
          lty=1)

# save the conflation coordinates into an object without plotting
# then we can plot or use the xy values if desired
R> conflated_normals <- conflate(c("density_fun(x, 'dnorm', mean=0, sd=1)",
                                   "density_fun(x, 'dnorm', mean=3, sd=1)"),
                                 from=-4, to=4, n=101, plot=FALSE)
R> plot(conflated_normals)
```

Conflation produces a useful way of summarizing information on the same parameter coming from different analyses. For instance, in the imaginary example above three normals were used to approximate the posterior distribution of a given divergence time (Figure 11), and their conflation is depicted as a black line. It is evident that the conflated distribution shows more resemblance with the component with lowest variance (blue distribution), although incorporating the information coming from all of them. This distribution can in turn be used as a secondary calibration point in a subsequent divergence time estimation study, where a single node calibration must be used for a given node.

The function `conflate` is used for building a conflation given arbitrary PDFs available in, that is, any function `dDIST` (e.g., `dnorm`, `dexp`, etc.). It uses a constructor function `density_fun` that helps to specify the functions without need to calculate any object beforehand.

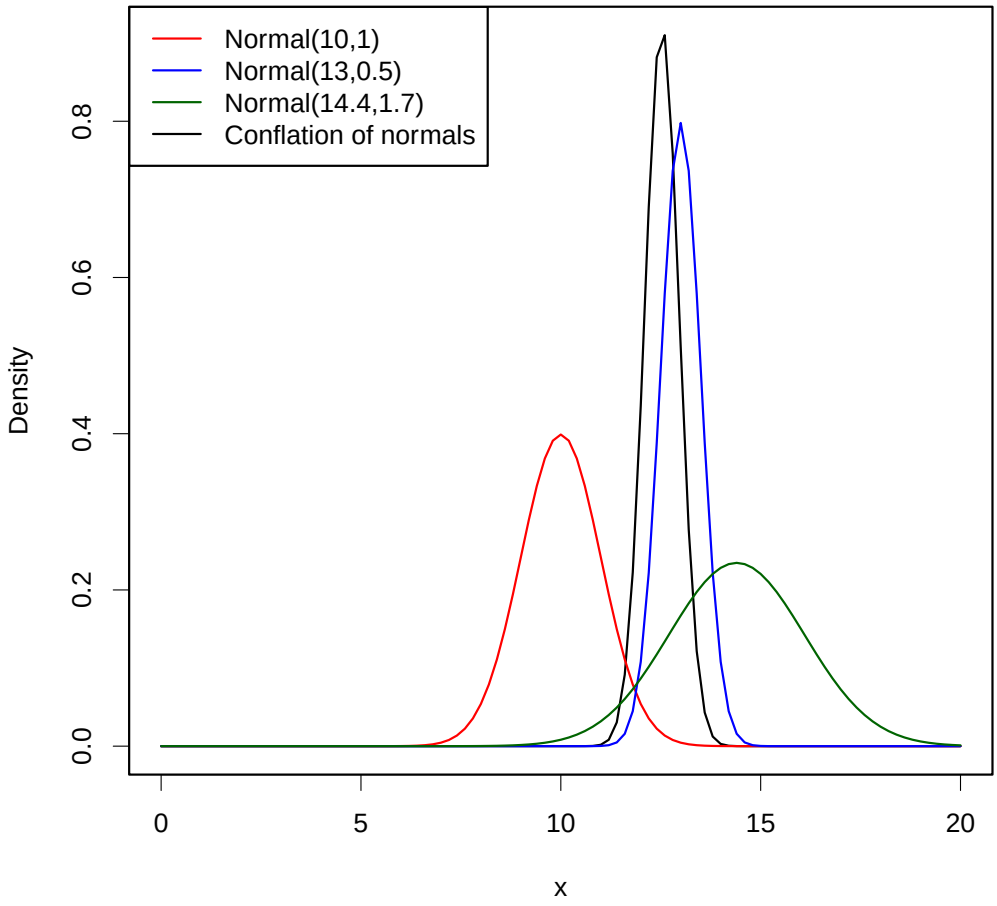

Figure 11: Conflation of three distributions representing imaginary posterior densities approximated by normals of the same divergence time in Ma (red, blue, and dark green lines). The conflated distribution of these three represents the information coming from each proportional to their variance (black line).

5 Discussion

Bayesian phylogenetic methods have had a dramatic impact in evolutionary biology. This is due in part to its ability to integrate multiple sources of information and to provide a natural way of characterizing uncertainty in parameter estimation. One of the challenges that face most analyses of empirical data is the reproducibility, although some programs have internal tools for improving this aspect of research. For instance, even though model specification is often carried out through the GUI program beauti, Beast2 will return the complete XML specification of the analysis as commented header of the log trace containing the parameter values from the MCMC sampling. The downside of this strategy is that the XML specification is very challenging to understand by readers (although convenient from a development point of view). Both MrBayes and RevBayes support interactive sessions so that the models can be specified piece by piece. However, their highest reproducibility standard is reached through the use of scripts, where the complete model is documented in code. MCMCTree is a special case where interactive mode is restricted to entering the path to the control file. The complete analysis needs to be specified in the control file and therefore, it qualifies as a script for practical purposes. The consequence is that any analysis specified in scripts is easier to understand and reproduce. Nonetheless, all these strategies are unable to document how the users chose the specific models, their parameters, and their distributions (e.g. priors).

The pre-analysis tools in the present package aim at allowing the user not just find the best specification of distributions and model components such as the starting tree, but also to document them as R scripts which can be shared alongside the analysis scripts. The same applies for post-analysis tools such as the comparison between model priors and posteriors, which are often verbally mentioned to have been done but not represented graphically. Finally, the tools for estimating events in time from collections of point estimates, as well as the combination of different posterior densities from the same parameter are expected to improve the transparency of research choices when e.g. justifying the choice of secondary calibration points, or discussing the timing of biogeographic events when multiple sources are available for them. An additional advantage of this toolset is to integrate information from multiple areas of biological and earth sciences. They allow to make a better use of geological and paleontological information for the specification of priors, as well as for estimating events in time, as some of the models (e.g. stratigraphic intervals) have been transferred here from paleobiology to the analysis of Bayesian estimates. Distribution conflation or estimation of parameters in stratigraphic intervals using estimates from divergence time estimation analyses, in turn, can be helpful for research in paleogeography and tectonics as they provide an independent source to contrast the predictions made with aid of geological tools. In the end, both areas are concerned with events in geologic time.

5.1 Comparison with other packages

Several packages have been developed for phylogenetics and related areas (Gearty et al., 2024). There are as many standards as software tools, and this is reflected in the multiplicity of approaches to software development applied to evolutionary biology. Multiple functionalities are also implemented in different ways, thus providing a high amount of flexibility to users of phylogenetic methods. Some tools available in this package are already available in a different form in some other packages (e.g. tools for interacting with data matrices, concatenation, and format conversion), while others are completely new to our knowledge (e.g. parameter estimation in distributions with a set of quantiles, confidence intervals on the x-intercept, or the conflation method). We think that having multiple implementations of the same tools is positive rather than redundant as much as the different implementations allow the same tools to be used in different analytical contexts rather than forcing the users to have the same settings, formats, and input.

`EvoPhylo` (Simões et al., 2023) targets divergence time estimation analyses using morphological datasets and their clock models. The package also pre- and post-process output from MrBayes but restricted to situations where morphological data, in particular for completely extinct taxon sampling, is present in the analysis. The package `beastier` (formerly `babette`; Bilderbeek and Etienne, 2008)

on the other hand, has code for interacting with programs of the `Beast` suite. Although some of our functions can be used to help specify analyses in `Beast2` and parse output logs from the program, our package is not intended to be a wrapper nor a gateway to `Beast`, but rather a tool for improving reproducibility when using that program. `CladeDate` (Claramunt, 2022) is very similar in that it aims at generating calibration densities from the fossil record. However, it differs in that package uses data argumentation via sampling from the uncertainty in fossil occurrences in order to generate a distribution of a parameter θ (fitted by using also methods for stratigraphic intervals) representing the origination of the clade. `phruta` (Román-Palacios, 2023) has tools for retrieving data from GenBank and then concatenate them for phylogenetic analysis. In contrast, the present package allows to concatenate matrices including both DNA and morphological data, so that total-evidence divergence time estimation and tree inference can be carried out combining these two data sources. The package `mcmc3r` (dos Reis et al., 2018; Álvarez-Carretero et al., 2019) targets analysis such as model selection using Bayes factors specifically using the program `MCMCTree`, whereas tools in `tbea` can be used for comparing results form independent runs using `MCMCTree`. Also, `mcmc3r` has functions for plotting the minimum and both upper and lower bound calibration densities used by `MCMCTree`, in a similar way as our package aids in plotting `Beast2`'s lognormal calibration density. Finally, `TreeTools` (Smith, 2019) implements tools for interacting with output from `TNT`. Although `tbea` has a function for reading trees in TNT format, its aim is to convert them to newick format rather than to generate an R representation of the tree; therefore, the operations on the TNT tree text representation return also a tree in text representation, but in newick format instead.

5.2 Dependencies

As mentioned earlier, one of our design choices was to reduce to a minimum possible the dependencies of `tbea`, which at the moment depends on `ape` (Paradis and Schliep, 2019) for phylogenetic tree representation and manipulation, `Rfit` (Kloke and McKean, 2012) for extending the possible linear models to be used when inferring the x-intercept of CDFs, and `coda` (Plummer et al., 2006) for calculating the highest posterior density interval which is used in crossplots.

6 Acknowledgments

This work was funded by FAPESP processes 2014/11558-5, 2016/02253-1, and 2023/07838-1 to GAB and 2020/09442-0, 2020/10206-9 and 2022/01533-1 to SR. We thank the feedback by students at the Universidad de Antioquia in Colombia during a graduate course on phylogenetics in 2021, and help by Klauss Schliep when implementing metaprogramming code for numerical optimization. We thank both Uwe Ligges and Benjamin Altmann for their feedback during submission to CRAN which helped improve the documentation of the package.

References

Abe, K. T., Mariguela, T. C., Avelino, G. S., Foresti, F., and Oliveira, C. (2014). Systematic and historical biogeography of the Bryconidae (Ostariophysi: Characiformes) suggesting a new rearrangement of its genera and an old origin of Mesoamerican ichthyofauna. *BMC evolutionary biology*, 14(1):152.

Álvarez-Carretero, S., Goswami, A., Yang, Z., and dos Reis, M. (2019). Bayesian estimation of species divergence times using correlated quantitative characters. *Systematic Biology*, 68(6):967–986.

Álvarez-Carretero, S., Tamuri, A. U., Battini, M., Nascimento, F. F., Carlisle, E., Asher, R. J., Yang, Z., Donoghue, P. C., and dos Reis, M. (2022). A species-level timeline of mammal evolution integrating phylogenomic data. *Nature*, 602(7896):263–267.

Antoine, P.-O., Abello, M. A., Adnet, S., Altamirano Sierra, A. J., Baby, P., Billet, G., Boivin, M., Calderón, Y., Candela, A., Chabain, J., Corfu, F., Croft, D. A., Ganerød, M., Jaramillo, C. A., Klaus, S., Marivaux, L., Navarrete, R. E., Orliac, M. J., Parra, F., Pérez, M. E., Pujos, F., Rage, J. C., Ravel, A., Robinet, C., Roddaz, M., Tejada-Lara, J. V., Vélez-Juarbe, J., Wesselingh, F. P., and Salas-Gismondi, R. (2016). A 60-million-year Cenozoic history of western Amazonian ecosystems in Contamana, eastern Peru. *Gondwana Research*, 31:30–59.

Ballen, G. A., Moreno-Bernal, J. W., and Jaramillo, C. (2022). The fossil record of sabre-tooth characins (Teleostei: Characiformes: Cynodontinae), their phylogenetic relationships and palaeobiogeographical implications. *Journal of Systematic Palaeontology*, 19(24):1679–1692.

Barba-Montoya, J., Tao, Q., and Kumar, S. (2021). Molecular and morphological clocks for estimating evolutionary divergence times. *BMC Ecology and Evolution*, 21(83):1–15.

Bilderbeek, R. J. and Etienne, R. S. (2008). babette: Beauti 2, beast 2 and tracer for r. *Methods in Ecology and Evolution*.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., Matschiner, M., Mendes, F. K., Müller, N. F., Ogilvie, H. A., du Plessis, L., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M. A., Wu, C.-H., Xie, D., Zhang, C., Stadler, T., and Drummond, A. J. (2019). Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *PLOS Computational Biology*, 15(4):1–28.

Cheviron, Z. A., Hackett, S. J., and Capparella, A. P. (2005). Complex evolutionary history of a Neotropical lowland forest bird (*Lepidothrix coronata*) and its implications for historical hypotheses of the origin of Neotropical avian diversity. *Molecular Phylogenetics and Evolution*, 36(2):338–357.

Claramunt, S. (2022). Cladedate: Calibration information generator for divergence time estimation. *Methods in Ecology and Evolution*, 13(11):2331–2338.

Collins, A. C. and Dubach, J. M. (2000). Biogeographic and ecological forces responsible for speciation in *Ateles*. *International Journal of Primatology*, 21(3):421–444.

Cortés-Ortiz, L., Bermingham, E., Rico, C., Rodríguez-Luna, E., Sampaio, I., and Ruiz-García, M. (2003). Molecular systematics and biogeography of the Neotropical monkey genus, *Alouatta*. *Molecular Phylogenetics and Evolution*, 26(1):64–81.

D'Horta, F. M., Cuervo, A. M., Ribas, C. C., Brumfield, R. T., and Miyaki, C. Y. (2013). Phylogeny and comparative phylogeography of *Sclerurus* (Aves: Furnariidae) reveal constant and cryptic diversification in an old radiation of rain forest understorey specialists. *Journal of Biogeography*, 40(1):37–49.

Dick, C. W., Abdul-Salim, K., and Bermingham, E. (2003). Diversification of a Widespread Tropical Rain Forest Tree. *The American Naturalist*, 162(6):691–703.

dos Reis, M., Gunnell, G. F., Barba-Montoya, J., Wilkins, A., Yang, Z., and Yoder, A. D. (2018). Using phylogenomic data to explore the effects of relaxed clocks and calibration strategies on divergence time estimation: Primates as a test case. *Systematic Biology*, 67(4):594–615.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*. Wiley Interscience, New York.

Drummond, A. J. and Stadler, T. (2016). Bayesian phylogenetic estimation of fossil ages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1699):20150129.

Elias, M., Joron, M., Willmott, K., Silva-Brandão, K. L., Kaiser, V., Arias, C. F., Piñerez, L. M. G., Uribe, S., Brower, A. V. Z., Freitas, A. V. L., and Jiggins, C. D. (2009). Out of the Andes: patterns of diversification in clearwing butterflies. *Molecular Ecology*, 18(8):1716–1729.

Fernandes, A. M., Wink, M., Sardelli, C. H., and Aleixo, A. (2014). Multiple speciation across the Andes and throughout Amazonia: The case of the spot-backed antbird species complex (*Hylophylax naevius/Hylophylax naevioides*). *Journal of Biogeography*, 41(6):1094–1104.

Gavryushkina, A., Welch, D., Stadler, T., and Drummond, A. J. (2014). Bayesian Inference of Sampled Ancestor Trees for Epidemiology and Fossil Calibration. *PLoS Computational Biology*, 10(12):e1003919.

Gearty, W., O'Meara, B., Berv, J., Ballen, G. A., Ferreira, D., Lapp, H., Schmitz, L., Smith, M. R., Upham, N. S., and Nations, J. A. (2024). Cran task view: Phylogenetics.

Genest, C. and Zidek, J. V. (1986). Combining Probability Distributions: A Critique and an Annotated Bibliography. *Statistical Science*, 1(1):114–135.

Goloboff, P. A., Farris, J. S., and Nixon, K. C. (2008). Tnt, a free program for phylogenetic analysis. *Cladistics*, 24(5):774–786.

Grau, E. T., Pereira, S. L., Silveira, L. F., Höfling, E., and Wajntal, A. (2005). Molecular phylogenetics and biogeography of Neotropical piping guans (Aves: Galliformes): Pipile Bonaparte, 1856 is synonym of Aburria Reichenbach, 1853. *Molecular Phylogenetics and Evolution*, 35(3):637–645.

Gutiérrez, E. E., Anderson, R. P., Voss, R. S., Ochoa-G., J., Aguilera, M., and Jansa, S. A. (2014). Phylogeography of *Marmosa robinsoni*: insights into the biogeography of dry forests in northern South America. *Journal of Mammalogy*, 95(6):1175–1188.

Hardman, M. and Lundberg, J. G. (2006). Molecular phylogeny and a chronology of diversification for "phractocephaline" catfishes (Siluriformes: Pimelodidae) based on mitochondrial DNA and nuclear recombination activating gene 2 sequences. *Molecular Phylogenetics and Evolution*, 40:410–418.

Heath, T. A., Huelsenbeck, J. P., and Stadler, T. (2014). The fossilized Birth-Death Process: A coherent model of fossil calibration for divergence time estimation. *Proceedings of the National Academy of Sciences*, 111(29):E2957–E2966.

Heled, J. and Drummond, A. J. (2009). Bayesian Inference of Species Trees from Multilocus Data. *Molecular Biology and Evolution*, 27(3):570–580.

Hernández Torres, C. L. (2015). *Species Delimitation, Phylogenetics, and Biogeography of the Catfish Genus* Rhamdia *Bleeker (Heptapteridae) of Central America and the Trans-Andean Region of Colombia*. Phd thesis, The University of Southern Mississippi.

Hill, T. P. (2008). Conflations of probability distributions. *Transactions of the American Mathematical Society*, 363:3351–3372.

Hill, T. P. (2011). Conflations of probability distributions. *Transactions of the American Mathematical Society*, 363(6):3351–3372.

Kendall, M. G. (1948). *Rank correlation methods*. Griffin.

Kloke, J. D. and McKean, J. W. (2012). Rfit: Rank-based estimation for linear models,. *The R Journal*, 4(2):57–64.

Kloke, J. D. and McKean, J. W. (2014). *Nonparametric Statistical Methods Using R*. CRC Press, 2nd edition.

Kreyszig, E. (1979). *Advanced Engineering Mathematics*. Wiley & Sons.

Landis, M. J. (2017). Biogeographic dating of speciation times using paleogeographically informed processes. *Systematic Biology*, 66(2):128–144.

Lewis, P. O. (2001). A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data. *Systematic Biology*, 50(6):913–925.

Machado, L. F., Leite, Y. L. R., Christoff, A. U., and Giugliano, L. G. (2014). Phylogeny and biogeography of tetralophodont rodents of the tribe Oryzomyini (Cricetidae: Sigmodontinae). *Zoologica Scripta*, 43(2):119–130.

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica*, 13(3):245–259.

Marshall, C. R. (1994). Confidence intervals on stratigraphic ranges: Partial relaxation of the assumption of randomly distributed fossil horizons. *Paleobiology*, 20(4):459–469.

Marshall, C. R. (2010). Using confidence intervals to quantify the uncertainty in the end-points of stratigraphic ranges. *Quantitative Methods in Paleobiology*, 16:291–316.

Miller, M. J., Bermingham, E., Klicka, J., Escalante, P., do Amaral, F. S. R., Weir, J. T., and Winker, K. (2008). Out of Amazonia again and again: episodic crossing of the Andes promotes diversification in a lowland forest flycatcher. *Proceedings of the Royal Society B*, 275(1639):1133–42.

Paradis, E. and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35:526–528.

Patané, J. S. L., Weckstein, J. D., Aleixo, A., and Bates, J. M. (2009). Evolutionary history of *Ramphastos* Toucans: Molecular phylogenetics, temporal diversification, and biogeography. *Molecular Phylogenetics and Evolution*, 53(3):923–934.

Patel, S., Weckstein, J. D., Patané, J. S. L., Bates, J. M., and Aleixo, A. (2011). Temporal and spatial diversification of *Pteroglossus* Araçaris (Aves: Ramphastidae) in the neotropics: Constant rate of diversification does not support an increase in radiation during the Pleistocene. *Molecular Phylogenetics and Evolution*, 58:105–115.

Picq, S., Alda, F., Krahe, R., and Bermingham, E. (2014). Miocene and Pliocene colonization of the Central American Isthmus by the weakly electric fish *Brachyhypopomus occidentalis* (Hypopomidae, Gymnotiformes). *Journal of Biogeography*, 41(8):1520–1532.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rannala, B. and Yang, Z. (2007). Inferring speciation times under an episodic molecular clock. *Systematic Biology*, 56(3):453–466.

Rannala, B. and Yang, Z. (2017). Efficient Bayesian Species Tree Inference under the Multispecies Coalescent. *Systematic Biology*, 66(5):823–842.

Ribas, C. C., Gaban-Lima, R., Miyaki, C. Y., and Cracraft, J. (2005). Historical biogeography and diversification within the Neotropical parrot genus *Pionopsitta* (Aves: Psittacidae). *Journal of Biogeography*, 32(8):1409–1427.

Ribas, C. C., Moyle, R. G., Miyaki, C. Y., and Cracraft, J. (2007). The assembly of montane biotas: linking Andean tectonics and climatic oscillations to independent regimes of diversification in *Pionus* parrots. *Proceedings. Biological sciences / The Royal Society*, 274(1624):2399–2408.

Říčan, O., Piálek, L., Zardoya, R., Doadrio, I., and Zrzavý, J. (2013). Biogeography of the Mesoamerican Cichlidae (Teleostei: Heroini): colonization through the GAARlandia land bridge and early diversification. *Journal of Biogeography*, 40(3):579–593.

Román-Palacios, C. (2023). The phruta r package: Increasing access, reproducibility and transparency in phylogenetic analyses. *Methods in Ecology and Evolution*, 14(9):2284–2299.

Ronquist, F., Klopfstein, S., Vilhelmsen, L., Schulmeister, S., Murray, D. L., and Rasnitsyn, A. P. (2012). A total-evidence approach to dating with fossils, applied to the early radiation of the hymenoptera. *Systematic Biology*, 61(6):973–999.

Ruiz-García, M., Vásquez, C., Sandoval, S., Kaston, F., Luengas-Villamil, K., and Shostell, J. M. (2015). Phylogeography and spatial structure of the lowland tapir (*Tapirus terrestris* , Perissodactyla: Tapiridae) in South America. *Mitochondrial DNA*, 1736(2013):1–9.

Sanderson, M. J. (2002). Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. *Molecular Biology and Evolution*, 19(1):101–109.

Seber, G. A. F. and Lee, A. J. (2013). *Linear Regression Analysis*. Wiley Interscience, Hoboken, NJ, 2nd editio edition.

Simões, T. R., Greifer, N., Barido-Sottani, J., and Pierce, S. E. (2023). Evophylo: An r package for pre- and postprocessing of morphological data from relaxed clock bayesian phylogenetics. *Methods in Ecology and Evolution*, 14(8):1981–1993.

Smith, B. T., Mccormack, J. E., Cuervo, A. M., Hickerson, M. J., Aleixo, A., Cadena, C. D., Pérez-Emán, J., Burney, C. W., Xie, X., Harvey, M. G., Faircloth, B. C., Glenn, T. C., Derryberry, E. P., Prejean, J., Fields, S., and Brumfield, R. T. (2014). The drivers of tropical speciation. *Nature*, pages 1–8.

Smith, M. R. (2019). *TreeTools: create, modify and analyse phylogenetic trees*. Comprehensive R Archive Network. R package version 1.11.0.

Smith, S. A. and O'Meara, B. C. (2012). treepl: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics*, 28(20):2689–2690.

Stadler, T. (2010). Sampling-through-time in birth–death trees. *Journal of Theoretical Biology*, 267(3):396–404.

Stadler, T. and Yang, Z. (2013). Dating phylogenies with sequentially sampled tips. *Systematic Biology*, 62(5):674–688.

Strauss, D. and Sadler, P. M. (1989). Classical confidence intervals and Bayesian probability estimates for ends of local taxon ranges. *Mathematical Geology*, 21(4):411–427.

To, T.-H., Jung, M., Lycett, S., and Gascuel, O. (2015). Fast dating using least-squares criteria and algorithms. *Systematic Biology*, 65(1):82–97.

Voss, R. S., Hubbard, C., and Jansa, S. A. (2013). Phylogenetic Relationships of New World Porcupines (Rodentia, Erethizontidae): Implications for Taxonomy, Morphological, and Biogeography. *American Museum Novitates*, (3769):1–36.

Warnock, R. C. M. and Wright, A. M. (2021). *Understanding the Tripartite Approach to Bayesian Divergence Time Estimation*. Elements of Paleontology. Cambridge University Press.

Weir, J. T. and Price, M. (2011). Andean uplift promotes lowland speciation through vicariance and dispersal in *Dendrocincla* woodcreepers. *Molecular Ecology*, 20(21):4550–4563.

Wickham, H. (2015). *R Packages*. O'Reilly Media.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Yang, Z. (2014). Models of nucleotide substitution. In *Molecular Evolution: A Statistical Approach*. Oxford University Press.