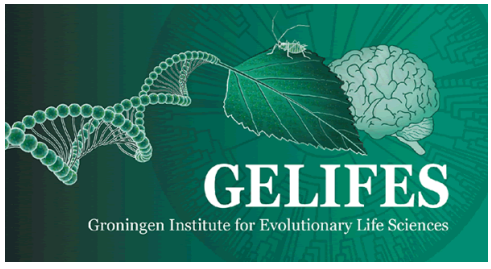
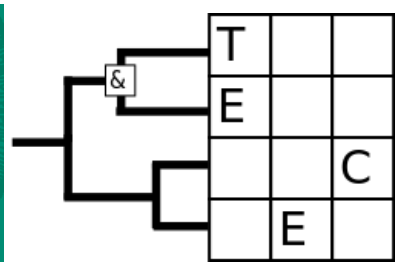
Speciation and the error we make in phylogenetic inference

Richèl J.C. Bilderbeek

university of groningen

faculty of science and engineering





Zernike Institute PhD thesis series 2020-09
ISSN: 1234-5678
ISBN: 123-45-678-9012-3
ISBN: 123-45-678-9012-3 (electronic version)

The work described in this thesis was performed in the research group Theoretical & Evolutionary Community Ecology at the University of Groningen, the Netherlands.

Cover design: Richèl J.C. Bilderbeek
Cover image: Common ash (*Fraxinus excelsior*), photo by Brian Green. This is the first tree shown at `https://en.wikipedia.org/wiki/Tree`.
An electronic version of this dissertation is available at:
`https://github.com/richelbilderbeek/thesis`.
Printed by: [name of printer]

university of groningen

Speciation and the error we make in phylogenetic inference

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. C. Wijmenga
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Friday September 18th 2020 at 16:15 hours

by

Richèl Jacobus Cornelis Bilderbeek

born on 2 September 1980
in Gennep, The Netherlands

Supervisor
Prof. R.S. Etienne

Co-supervisor
Dr. A.L. Pigot

Assessment Committee
Prof. F. Hartig
Prof. L. Harmon
Prof. E. Wit

CONTENTS

1

INTRODUCTION

1

Once upon a time, there was the evolution of all life on Earth. Let me tell the simplified version of this story and how to put this into figures called phylogenies, before moving to the more complex details. The formation of the Earth began approximately 4.5 billion years ago (Dalrymple 2001). From an evolutionary biologists' point of view, this was a dull time, until the first living organism appeared.

This First Universal Common Ancestor (FUCA) came into existence at least 3.48 billion years ago (Noffke *et al.* 2013). FUCA may not have been alone, but these other early life forms went extinct [1] and are ignored in this story . We can depict the evolutionary history of FUCA at that point in time with figure 1.1.

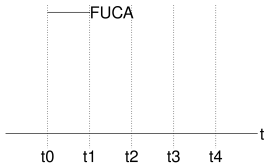


Figure 1.1 | Evolutionary history of the First Universal Common Ancestor (FUCA). Time goes from past (left) towards the present (right).

One unknown day, the descendants of FUCA became dissimilar enough to say that the one species called FUCA gave rise to two species (note the difficulty in determining what a species is at that time!). This event doubled the biodiversity on Earth. The two species that FUCA evolved into will be called species A and B. Species A and B are sister species. We can depict the evolutionary history of these two species in figure 1.2.

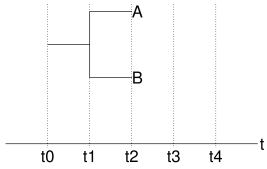


Figure 1.2 | Evolutionary history of the two descendants of FUCA. Time goes from past (left) towards the present (right).

Both species A and B have their unknown histories. One of them may have gone extinct, as extinction is a common event: it is estimated that more than 99% of all species that has ever lived on Earth has gone extinct (Newman 1997). Alternatively, they may have given rise to new species, but these are just as likely to go extinct. For this story, we will assume A and/or the clade of its descendant species went extinct and that species B created a sister species C. Species B and C gave rise to all contemporary biodiversity. This ancestor of species B and C is called the Last Universal Common Ancestor, or LUCA. LUCA is estimated to have lived between 3.48 (Noffke *et al.* 2013) and 4.5 (Betts *et al.* 2018) billions of years ago. We can depict the evolutionary history of LUCA in figure 1.3. Here, billions of years ago, is where the story ends and we will move on to the present.

[1] by definition!

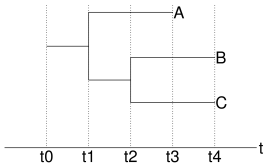


Figure 1.3 | Evolutionary history of the three descendants of FUCA, of which one (A) went extinct. Assuming B and C gave rise to all contemporary biodiversity, the Last Universal Common Ancestor (LUCA) existed at timepoint t2. Time goes from past (left) towards the present (right).

The idea that all life on Earth is related was first posed by Charles Darwin in his book 'On the Origin of Species' in 1859 (Darwin 1859). His first sketch of an evolutionary tree is shown in figure 1.4.

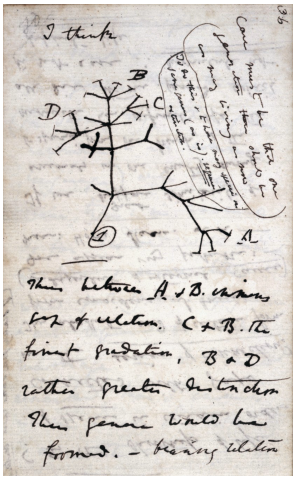


Figure 1.4 | Charles Darwin's first sketch of an evolutionary tree (1837).

The biodiversity derived from the first life on Earth is important to us humans (apart from that is has created us) for many reasons. One of these is that biodiversity usually improves ecosystem services (Cardinale *et al.* 2012), where ecosystem services are features of biological systems that are positive for human well-being, for example food, carbon sequestration, waste decomposition and pest control. Therefore, biodiversity is linked to human well-being. Biodiversity is considered so important that the European Union has an explicit Biodiversity Strategy, which aims to halt the loss of biodiversity (see `https://ec.europa.eu/environment/nature/biodiversity/strategy/index_en.htm`).

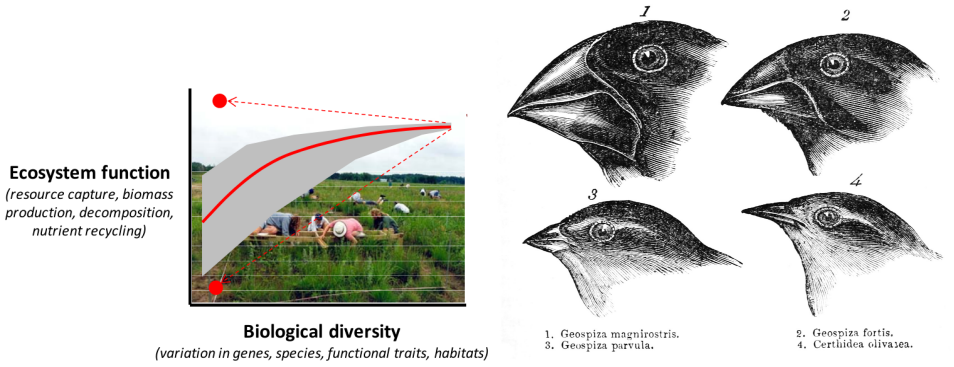


Figure 1.5 | Left: a diversity-function relationship found to be typical from hundreds of studies. The red line represents an average, where the gray polygon represents a 95% confidence interval. The red dots show the lower and upper limit for monocultures. From Cardinale *et al.* 2012. Right: Darwin's finches, by John Gould.

Speciation is the process that increases biological diversity. This process is studied from multiple angles; among others, we can study the mechanism ('what causes a speciation event?') or we can study the patterns of many of such events ('is speciation rate constant through time?'). Darwin's finches (see figure 1.5) represent an iconic example of speciation with 25,000 results on Google Scholar. There are many suggested mechanisms underlying speciation events, such as reproductive incompatibilities arising in geographical isolation (e.g. Mayr 1942), ecological factors (e.g. Lack 1947) causing divergent selection, and sexual selection resulting in assortative mating. However, listing and explaining all mechanisms is beyond the scope of this thesis. In this thesis I assume speciation occurs and I focus on the questions what impact it has on evolutionary relationships between species and how we can infer speciation events from observed evolutionary relationships, as encoded in a phylogenetic tree. Getting such a phylogeny is not trivial, as I will discuss below. But once we have such a phylogeny, we can ask many question such as 'How often do speciation and extinction events take place?' 'Are speciation and extinctions rates constant, or do they change?', 'What causes a change in the speciation rate or the extinction rate?' or 'Is there an upper limit to the number of species?'.

There are two methods to study speciation patterns in evolutionary time: the use of fossils or the use of molecular phylogenies.

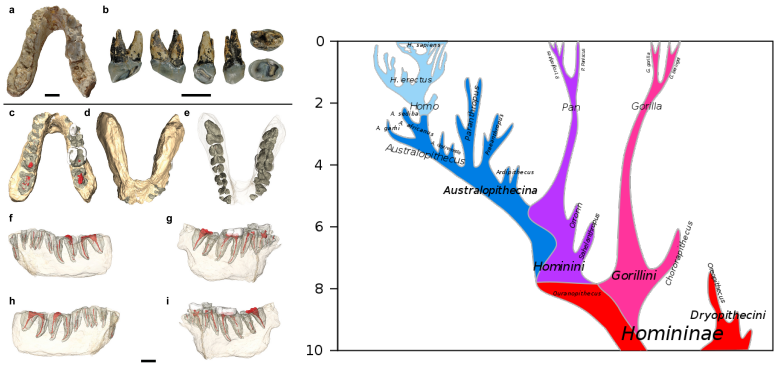


Figure 1.6 | Left: El Graeco fossil, from Fuss *et al.* 2017. Right: Evolution of the Homininae, based on Stringer 2012

Using fossils is a classic way to look back in evolutionary time. Fossils show a glimpse of the biodiversity in the past. We can deduce the age of fossils, by dating the rock layers they are found in. Using fossils has its limitations. First, it is mostly species with hard body parts that fossilize. Even in such species, organisms are only rarely preserved, and only a fraction of preserved fossils are preserved under ideal circumstances. Of these fossils, only a fraction is discovered. One example of a famous fossil is 'El Graeco', which may be the oldest known hominin (Fuss *et al.* 2017), where hominins are the tribe (taxonomic group) we Homo sapiens share with the Panini.

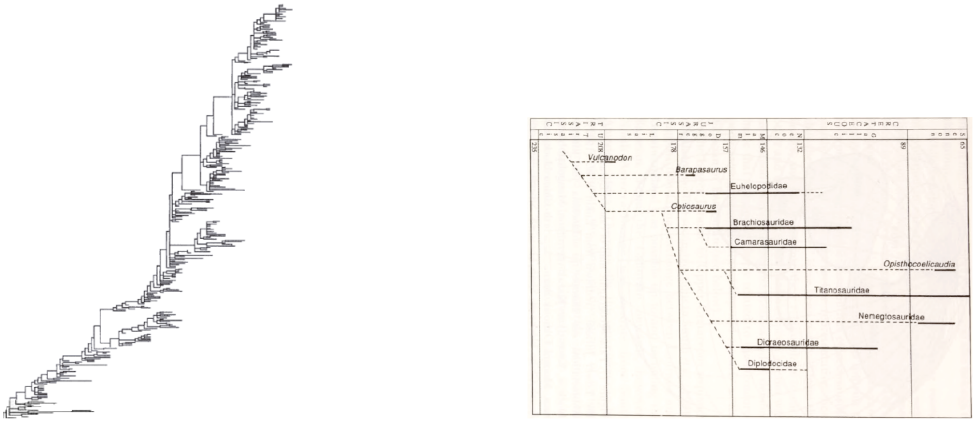


Figure 1.7 | Left: phylogeny of the human influenza virus type A subtype H3, from Bush *et al.* 1999. Right: the evolutionary history of sauropod dinosaurs, from Upchurch 1995

Using molecular phylogenies is the modern way to look back in evolutionary time. It is the use of heritable molecules (for example DNA, RNA, or proteins) of contemporary species to infer phylogenies. The field of phylogenetics is the research discipline that intends to infer the most accurate phylogenies possible, regarding topology, speciation and extinction times, optionally adding morphological data and/or fossil data. Phylogenetics

is applied in many settings, among others, species classification, forensics, conservation ecology and epidemiology (Lam *et al.* 2010).

One example of the importance of an accurate phylogenetic tree is demonstrated in Bush *et al.* 1999. This study investigated which loci of the H3 hemagglutinin surface protein are under selection, by contrasting nonsynonymous and synonymous mutation rates along the branches of a phylogeny. They noted that most selection rates were either below or above the statistical threshold depending on the phylogeny. This study contributed to recommendations on the composition of influenza virus vaccines.

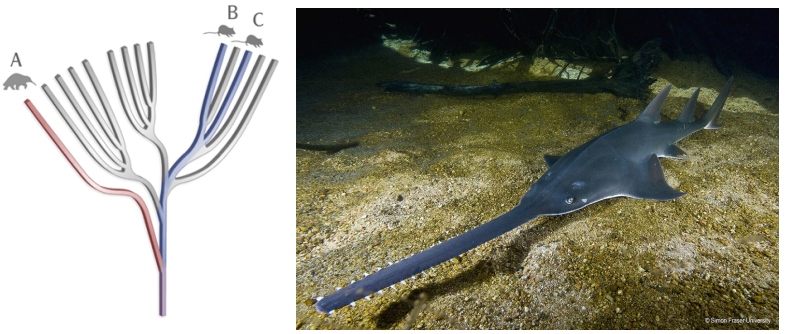


Figure 1.8 | Left: The ED (evolutionary distinctiveness) of species A is higher than that of species B or C, as more evolutionary history will be lost when that species goes extinct. Right: The Largetooth Sawfish (Pristis pristis) is at number 1 of the EDGE (ED = 'Evolutionary Distinctiveness', GE = Globally Endangered status) list, with an EDGE Score of 7.38 and an ED of 99.298.

Another example of the importance of an accurate phylogenetic tree comes from conservation biology, in which phylogenies are used to calculate an EDGE ('Evolutionarily Distinct and Globally Endangered') score. Species with a high EDGE score are prioritized in conservation. To calculate an EDGE score, one needs a metric of evolutionary distinctiveness ('ED') and globally 'endangeredness' ('GE'). The GE score is a conservational status, ranging from zero ('Least Concern') to four ('Critically Endangered'). The ED embodies the amount of evolutionary history lost if the species went go extinct, which can be calculated from a (hopefully accurate) phylogeny.

Phylogenetics has taken a huge flight, due to the massively increased computational power and techniques. A first milestone in this field is the work of Felsenstein in 1980, creating (and still maintaining!) PHYLIP (Felsenstein 1981), the first software package for classical phylogenetic analysis. Another milestone is the Metropolis-Hasting algorithm, which allowed Bayesian phylogenetics to thrive, resulting in contemporary tools such as BEAST (Drummond & Rambaut 2007), BEAST2 (Bouckaert *et al.* 2019) (of which more below), MrBayes (Huelsenbeck & Ronquist 2001) and RevBayes (Höhna *et al.* 2016).

Figure 1.9 | Left: PHYLIP logo. Center: BEAST2 logo. Right: BEAST2 example output

A clear example of the power of modern phylogenetics, is the Tree Of Life. The Tree Of Life is based on the proteome of 3,083 species. A proteome of a species consists of all the proteins found within that species. To be able to compare between different species, the researchers used part of the proteome that is common in most of these species, which consisted of 2,596 amino-acids. To create the Tree of Life, it took 3,840 computational hours on a modern supercomputer (Hug *et al.* 2016).

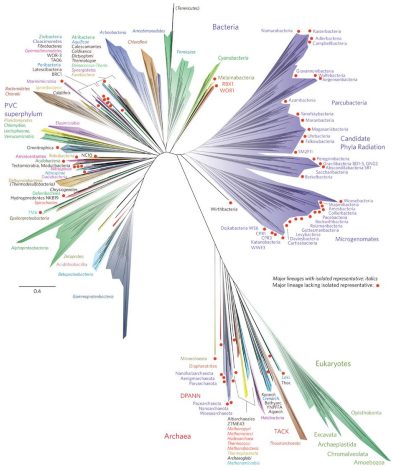


Figure 1.10 | Tree of Life, from Hug *et al.* 2016

To create such a tree from protein sequences, one has to specify an evolutionary model. This evolutionary model embodies our set of assumptions, such as the way a protein sequence evolves (also called the site model), the rate(s) at which this happens (the clock model) and the way in which a branching/speciation event takes place (the tree model). For example, the amino acids of the Tree Of Life are assumed to change over time according to the LG model (Le & Gascuel 2008). The speeds at which amino acids change to others are called the transition rates. The LG model is a model for amino acid transitions, which uses the average rates found in nature.

There are many evolutionary models to choose from, and selecting which one to use is hard, due to the many sets of assumptions to choose from. In general, modelers are looking for that set of assumptions that is as simple as possible, but not simpler. And even then, sometimes an overly simplistic model is still picked, due to computational

constraints.

Ideally, one would like to have a rational way to select an evolutionary model that is as simple as possible, but not simpler. Model comparison algorithms have been developed that select the evolutionary model that is most likely to have generated the data, without being overly complex. The idea is that the best evolutionary model should result in the most accurate phylogenetic trees.

Because model comparison is hard, there have been multiple studies investigating the effect of picking the wrong evolutionary models.

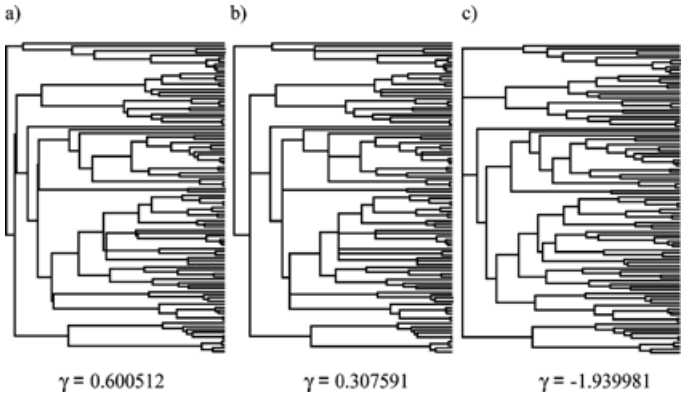


Figure 1.11 | Figure from Revell *et al.* 2005. Left: true tree. Middle: inferred tree, inferred using the generative model (i.e. the model that generated the true tree) Right: inferred tree, inferred using an inference model that is simpler than the generative model

One example that demonstrates the effect of using a too simple inference model is provided by Revell and colleagues (Revell *et al.* 2005). They first simulated many phylogenies. From those phylogenies, they simulated DNA sequences for each of the virtual species. DNA is the heritable material all life on Earth possesses, which consists of a sequence of the four DNA nucleotides. In the simulation of the DNA sequences, the experimenters used different DNA substitution models. A DNA substitution model embodies the transition rates of these nucleotides (see figure 1.19 for an example). From the simulated DNA sequences, the researchers inferred phylogenies again, with either the correct or a simpler DNA substitution model. Ideally, the inferred phylogenies match the phylogenies the alignments are based upon. They found that when the DNA model is the correct one, inference of the phylogenies is not flawless but satisfactory. However, when using an overly simplistic DNA model, the inferred tree shows a slowdown in their speciation rates, even when the original tree was simulated with a constant speciation rate. This study shows that a decreasing speciation rate may be attributed to an overly simplistic DNA model, instead of an interesting biological process.

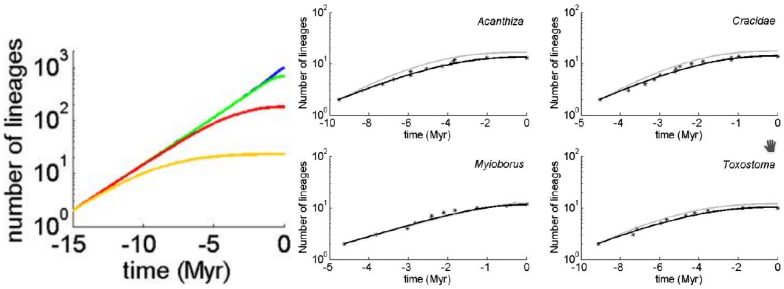


Figure 1.12 | Left: example lineage-through-time plots, for different speciation completion rates: yellow = 0.01, red = 0.1, green = 1.0, blue = 10. Note the slowdown in the accumulation of new lineages when speciation completion rate is lowered. Right: number of species through time plots for four bird phylogenies, (after Phillimore & Price 2008) Both figures are adapted from Etienne & Rosindell 2012

A more recent example that demonstrates the effect of using an overly simple inference model is the study by Duchêne and co-workers (Duchêne *et al.* 2014), who looked into the consequences of assuming a wrong clock model. A clock model embodies our assumptions regarding the mutation rates in the histories of different taxa. The simplest clock model, called the strict clock model, assumes these mutation rates are equal across all taxa. Using a wrong clock model has a profound impact on the inferred phylogenetic trees, unless we can specify the timing of some early speciation events (Duchêne *et al.* 2014).

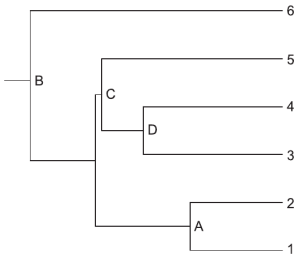


Figure 1.13 | Phylogeny with speciation events labelled A to D, where B is the earliest speciation event. Figure from Duchêne *et al.* 2014.

The tree model is the most important part of the evolutionary model needed for phylogenetic inference, with regard to speciation. The assumptions of a tree model are collectively called the tree prior, where 'prior' refers to the knowledge known before creating a phylogeny. The tree prior specifies the probability of processes that determine the shape of a tree. These two processes are (1) the formation of a new branch, and (2) the termination of an existing branch. In the context of speciation, we call these two events a speciation and an extinction event respectively.

There are two standard tree models, called the Yule and Birth-Death model. The most basic speciation model is the Yule model (Yule 1925), which assumes that speciation is constant and there is no extinction. Although extinction is a well-established phenomenon, the utility of the Yule model is its simplicity: it is the simplest evolutionary

model to work with, and the computation of the probability of a tree under the Yule process is very fast, making it a good first step in an evolutionary experiment. Similar to the simplest models of bacterial growth, the Yule model predicts that the expected number of species grows exponentially through time. Because the Yule model was later classified as a Birth-Death model without extinction, it is nowadays also called the Pure-Birth model.

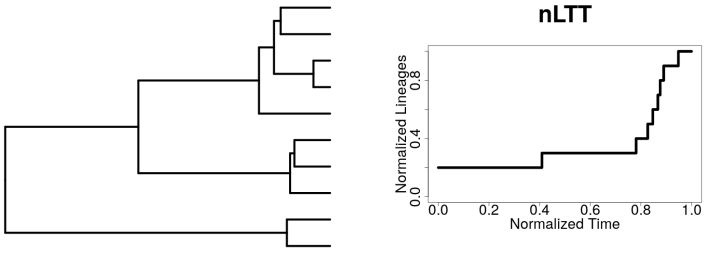


Figure 1.14 | Left: An example Yule tree Right: A lineages-through-time plot of the example Yule tree. In all cases, time goes from past (left) towards the present (right).

The Birth-Death model (Nee S., May R. M. & Harvey P. H. 1994) is an extension of the Yule model, as it adds extinction. Similar to the constant birth rate, the extinction rate is assumed to be constant as well. As a consequence, the BD model predicts two outcomes: if the speciation rate exceeds the extinction rate, the expected number of extant species grows exponentially through time. The other way around, however, when the extinction rate exceeds the speciation rate, the expected number of lineages is expected to decline exponentially. It is clear that exponential growth in the expected number of lineages is biologically nonsensical. To state the obvious: a finite area (Earth) results in a finite number of species. Applying the BD model to molecular data already shows that it does not always hold, as shown by figure 1.15.

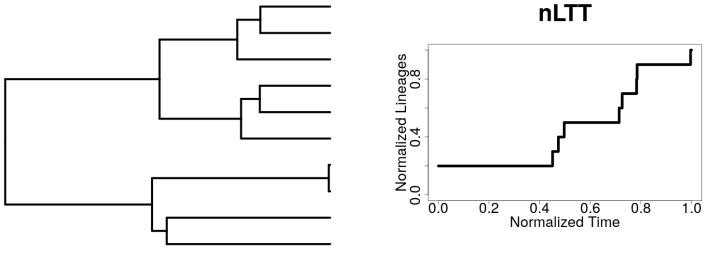


Figure 1.15 | Left: An example Birth-Death tree Right: A lineages-through-time plot of the example Birth-Death tree. In all cases, time goes from past (left) towards the present (right).

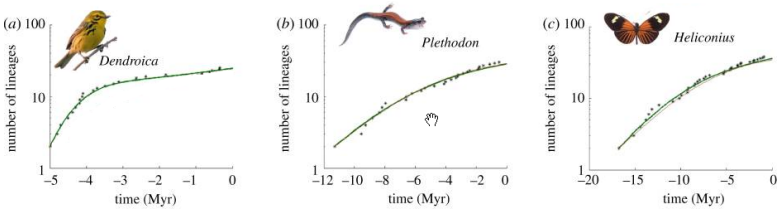


Figure 1.16 | An LTT plot for bird/lizards showing a slowdown in speciation rate, adapted from Etienne *et al.* 2012. Because the number of lineages on the y-axis are plotted on a logarithmic scale, exponential growth would show as a straight line.

A recent study investigating the effect of picking a wrong standard tree prior was provided by Sarver and colleagues (Sarver *et al.* 2019). In this study, they first simulated trees using either a Yule or a birth-death tree model, after which they simulated an alignment from that phylogeny using two different standard clock models. From these alignments, they inferred the original trees using all of the four different clock and tree prior combinations. They showed that, regardless of which priors are used, the estimated speciation and diversification rates from the inferred trees are similar to those of the original tree.

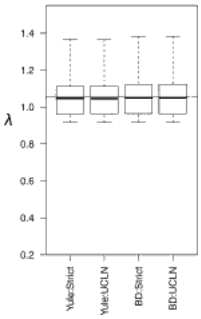


Figure 1.17 | Estimation of the speciation rate (λ) on inferred trees using 4 evolutionary models. The original trees had 100 taxa and were simulated with a strict clock model and BD tree model, with a speciation rate of 1.104. Adapted from Sarver *et al.* 2019.

This thesis investigates the effect of picking a wrong standard tree prior, when the tree is generated by a non-standard, novel tree model. I will describe one new biologically relevant tree model, as well as the re-usable framework to determine the effect of using a standard tree prior.

This novel and non-standard tree model is the multiple-birth death (MBD) model by Laudanno and colleagues (unpublished). While the standard BD models assume that a speciation event occurs in one species only at a time, the MBD models allows for speciation events to occur in multiple species at the same time. The biological idea behind this model, is that when a habitat (lake or mountain range) gets split into two, this may trigger speciation events in both communities at the same time. This mechanism is posed as an explanation for high biodiversity in the African rift lake Tanganyika, where

1

the water level rises and falls with ice ages, splitting up and merging the lake again and again, triggering co-occurring speciation events at each change.

This thesis investigates the effect of picking a wrong standard tree prior, when the tree is generated by a non-standard tree model, using the phylogenetic software called BEAST2 (Bouckaert *et al.* 2019), an abbreviation of 'Bayesian Evolutionary Analysis by Sampling Trees'.

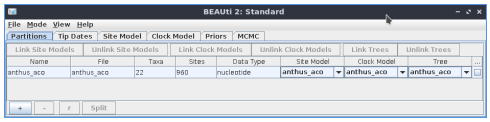


Figure 1.18 | BEAUti, after having picked a DNA alignment

We chose to use BEAST2 (Bouckaert *et al.* 2019) over other phylogenetic software, because BEAST2 is popular, beginner-friendly, flexible, has a package manager and a modular well-designed software architecture. The beginner-friendliness comes from the BEAST2 program called BEAUti, in which the user can set up his/her evolutionary model from a graphical user interface. There are many (in the order of dozens to hundreds) options to set up an evolutionary inference model. These choices are categorized in a site model, clock model and a tree prior.

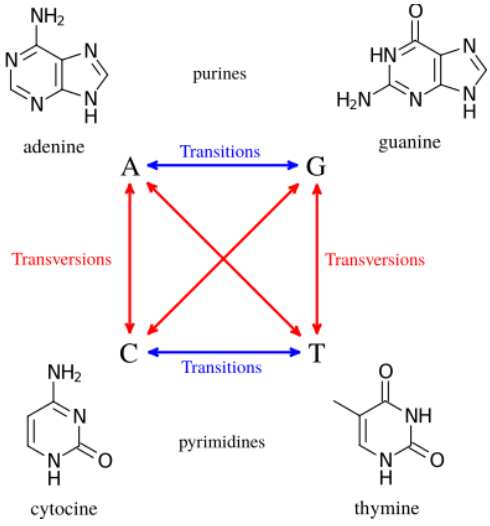


Figure 1.19 | Classification of nucleotide substitutions. The simplest nucleotide substitution model (JC69) assumes all 6 rates are equal, whereas the most complex one (GTR) allows all of these to differ.

A site model embodies the way the characters - nucleotides in our case of DNA sequences - change over time. One can specify the proportion of nucleotides that changes, or let it be estimated. Furthermore, one can specify how dissimilar different transition rates may be between different nucleotides. Most essential is the nucleotide substitution

model, which entails the relation between the twelve transition rates from any of the four nucleotides to any of the other three nucleotides. The simplest model (called JC69) assumes all are equal, whereas the most complex model (called GTR) assumes that all may differ. The standard BEAST2 software has four site models, but there is a BEAST2 package that contains 18 additional nucleotide substitution models.

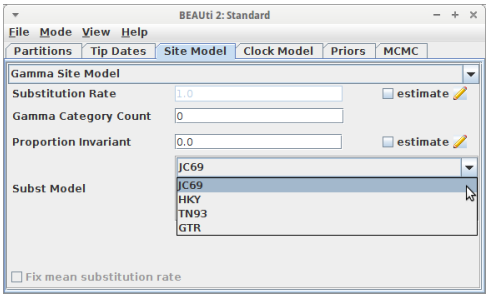


Figure 1.20 | Specifying a site model in BEAUti

To give an idea of the flexibility of BEAST2, I will zoom in on specifying one simple aspect of the inference model: the proportion invariants. The proportion invariants is the proportion, ranging from a value of zero (for 'none') to one (for 'all'), of nucleotides that remains unchanged throughout the evolutionary history. This proportion can either be set to a certain value, or be estimated. If the value is set to a certain value, BEAST2 assumes this as the truth. If the value is to be estimated by BEAST2, then one must additionally specify an initial value and a distribution how probable the different values are. By default, BEAST2 assumes a uniform distribution, that assigns an equal probability to all values between (and including) zero and one. Instead of using a uniform distribution, there are ten other distributions that can be picked as well, allowing, for example, to assign higher probabilities to certain proportions. So, for one simple value, there is already a plethora of options, and there are even more that I will not discuss. Within BEAST2, this liberty is the rule, instead of the exception, rendering it very flexible.

The clock model embodies how the mutation rates vary between different species. The simplest clock model, called the strict clock, assumes that mutation rates are identical in all species at all times. Two models (called relaxed-clock models) assume that mutation rates between species are independent (yet all rates are from one probability distribution), but stay constant after each species' inception. The last standard clock model (called a random local clock) assumes that all species have the same mutation rate at any time, yet the mutation rates varies through time.

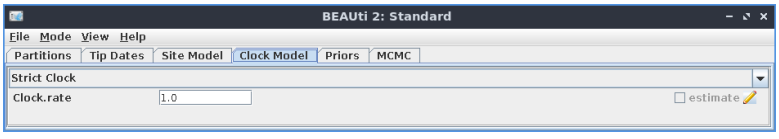


Figure 1.21 | Specifying a clock model in BEAUti

1

The tree prior specifies how a tree is built up, or, in our context, how speciation takes place in time, at the macro-evolutionary level. In our context, these are the Yule and Birth-Death model, which I already described earlier.

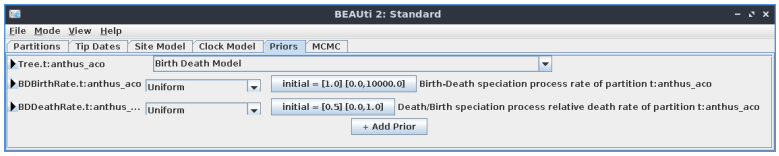


Figure 1.22 | Specifying a tree prior in BEAUti

This thesis investigates the effect of picking a wrong standard tree prior, when the tree is generated by a non-standard tree model. It does so, by using the same experimental setup, called 'pirouette', which is described in chapter 3. This framework is built up a foundation of R packages called 'babette', which is described in chapter 2.

Figure 1.23 | Environment that follows an unknown speciation model.

In the end, we want to know how well we can infer a phylogeny from molecular data found in the field. That field, outside, follows an unknown speciation model. Rather than just hope that our inference is robust to whatever novel model we throw at it, with this thesis I have aimed at providing methodology that can assess that robustness.

REFERENCES

Betts, H.C., Puttick, M.N., Clark, J.W., Williams, T.A., Donoghue, P.C. & Pisani, D. (2018) Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nature ecology & evolution*, **2**, 1556.

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. & Drummond, A.J. (2019) Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *PLOS Computational Biology*, **15**, 1–28.

Bush, R.M., Fitch, W.M., Bender, C.A. & Cox, N.J. (1999) Positive selection on the h3 hemagglutinin gene of human influenza virus a. *Molecular biology and evolution*, **16**, 1457–1465.

Cardinale, B.J., Duffy, J.E., Gonzalez, A., Hooper, D.U., Perrings, C., Venail, P., Narwani, A., Mace, G.M., Tilman, D., Wardle, D.A. *et al.* (2012) Biodiversity loss and its impact on humanity. *Nature*, **486**, 59–67.

Dalrymple, G.B. (2001) The age of the earth in the twentieth century: a problem (mostly) solved. *Geological Society, London, Special Publications*, **190**, 205–221.

Darwin, C. (1859) On the origin of species.

Drummond, A.J. & Rambaut, A. (2007) Beast: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, **7**, 214.

Duchêne, S., Lanfear, R. & Ho, S.Y. (2014) The impact of calibration and clock-model choice on molecular estimates of divergence times. *Molecular phylogenetics and evolution*, **78**, 277–289.

Etienne, R.S., Haegeman, B., Stadler, T., Aze, T., Pearson, P.N., Purvis, A. & Phillimore, A.B. (2012) Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 1300–1309.

Etienne, R.S. & Rosindell, J. (2012) Prolonging the past counteracts the pull of the present: protracted speciation can explain observed slowdowns in diversification. *Systematic Biology*, **61**, 204.

Felsenstein, J. (1981) Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, **17**, 368–376.

Fuss, J., Spassov, N., Begun, D.R. & Böhme, M. (2017) Potential hominin affinities of graecopithecus from the late miocene of europe. *PloS one*, **12**, e0177127.

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

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

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K. *et al.* (2016) A new view of the tree of life. *Nature microbiology*, **1**, 16048.

Lack, D. (1947) The significance of clutch-size. *Ibis*, **89**, 302–352.

Lam, T.T.Y., Hon, C.C. & Tang, J.W. (2010) Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections. *Critical reviews in clinical laboratory sciences*, **47**, 5–49.

Le, S.Q. & Gascuel, O. (2008) An improved general amino acid replacement matrix. *Molecular biology and evolution*, **25**, 1307–1320.

Mayr, E. (1942) *Systematics and the origin of species, from the viewpoint of a zoologist*.

Nee S., May R. M. & Harvey P. H. (1994) The reconstructed evolutionary process. *Phil Trans R Soc Lond B*, **344**, 305–311.

Newman, M.E.J. (1997) A model of mass extinction.

Noffke, N., Christian, D., Wacey, D. & Hazen, R.M. (2013) Microbially induced sedimentary structures recording an ancient ecosystem in the ca. 3.48 billion-year-old dresser formation, pilbara, western australia. *Astrobiology*, **13**, 1103–1124.

Phillimore, A.B. & Price, T.D. (2008) Density-dependent cladogenesis in birds. *PLoS biology*, **6**.

Revell, L.J., Harmon, L.J. & Glor, R.E. (2005) Under-parameterized model of sequence evolution leads to bias in the estimation of diversification rates from molecular phylogenies. *Systematic Biology*, **54**, 973–983.

Sarver, B.A., Pennell, M.W., Brown, J.W., Keeble, S., Hardwick, K.M., Sullivan, J. & Harmon, L.J. (2019) The choice of tree prior and molecular clock does not substantially affect phylogenetic inferences of diversification rates. *PeerJ*, **7**, e6334.

Stringer, C. (2012) What makes a modern human. *Nature*, **485**, 33–35.

Upchurch, P. (1995) The evolutionary history of sauropod dinosaurs. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, **349**, 365–390.

Yule, G.U. (1925) A mathematical theory of evolution, based on the conclusions of dr. jc willis, frs. *Philosophical transactions of the Royal Society of London Series B, containing papers of a biological character*, **213**, 21–87.

REFERENCES

Betts, H.C., Puttick, M.N., Clark, J.W., Williams, T.A., Donoghue, P.C. & Pisani, D. (2018) Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nature ecology & evolution*, **2**, 1556.

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. & Drummond, A.J. (2019) Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *PLOS Computational Biology*, **15**, 1–28.

Bush, R.M., Fitch, W.M., Bender, C.A. & Cox, N.J. (1999) Positive selection on the h3 hemagglutinin gene of human influenza virus a. *Molecular biology and evolution*, **16**, 1457–1465.

Cardinale, B.J., Duffy, J.E., Gonzalez, A., Hooper, D.U., Perrings, C., Venail, P., Narwani, A., Mace, G.M., Tilman, D., Wardle, D.A. *et al.* (2012) Biodiversity loss and its impact on humanity. *Nature*, **486**, 59–67.

Dalrymple, G.B. (2001) The age of the earth in the twentieth century: a problem (mostly) solved. *Geological Society, London, Special Publications*, **190**, 205–221.

Darwin, C. (1859) On the origin of species.

Drummond, A.J. & Rambaut, A. (2007) Beast: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, **7**, 214.

Duchêne, S., Lanfear, R. & Ho, S.Y. (2014) The impact of calibration and clock-model choice on molecular estimates of divergence times. *Molecular phylogenetics and evolution*, **78**, 277–289.

Etienne, R.S., Haegeman, B., Stadler, T., Aze, T., Pearson, P.N., Purvis, A. & Phillimore, A.B. (2012) Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 1300–1309.

Etienne, R.S. & Rosindell, J. (2012) Prolonging the past counteracts the pull of the present: protracted speciation can explain observed slowdowns in diversification. *Systematic Biology*, **61**, 204.

Felsenstein, J. (1981) Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, **17**, 368–376.

Fuss, J., Spassov, N., Begun, D.R. & Böhme, M. (2017) Potential hominin affinities of graecopithecus from the late miocene of europe. *PloS one*, **12**, e0177127.

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

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

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K. *et al.* (2016) A new view of the tree of life. *Nature microbiology*, **1**, 16048.

Lack, D. (1947) The significance of clutch-size. *Ibis*, **89**, 302–352.

Lam, T.T.Y., Hon, C.C. & Tang, J.W. (2010) Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections. *Critical reviews in clinical laboratory sciences*, **47**, 5–49.

Le, S.Q. & Gascuel, O. (2008) An improved general amino acid replacement matrix. *Molecular biology and evolution*, **25**, 1307–1320.

Mayr, E. (1942) *Systematics and the origin of species, from the viewpoint of a zoologist*.

Nee S., May R. M. & Harvey P. H. (1994) The reconstructed evolutionary process. *Phil Trans R Soc Lond B*, **344**, 305–311.

Newman, M.E.J. (1997) A model of mass extinction.

Noffke, N., Christian, D., Wacey, D. & Hazen, R.M. (2013) Microbially induced sedimentary structures recording an ancient ecosystem in the ca. 3.48 billion-year-old dresser formation, pilbara, western australia. *Astrobiology*, **13**, 1103–1124.

Phillimore, A.B. & Price, T.D. (2008) Density-dependent cladogenesis in birds. *PLoS biology*, **6**.

Pybus, O.G. & Harvey, P.H. (2000) Testing macro–evolutionary models using incomplete molecular phylogenies. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **267**, 2267–2272.

Revell, L.J., Harmon, L.J. & Glor, R.E. (2005) Under-parameterized model of sequence evolution leads to bias in the estimation of diversification rates from molecular phylogenies. *Systematic Biology*, **54**, 973–983.

Sarver, B.A., Pennell, M.W., Brown, J.W., Keeble, S., Hardwick, K.M., Sullivan, J. & Harmon, L.J. (2019) The choice of tree prior and molecular clock does not substantially affect phylogenetic inferences of diversification rates. *PeerJ*, **7**, e6334.

Stringer, C. (2012) What makes a modern human. *Nature*, **485**, 33–35.

Upchurch, P. (1995) The evolutionary history of sauropod dinosaurs. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, **349**, 365–390.

Yule, G.U. (1925) A mathematical theory of evolution, based on the conclusions of dr. jc willis, frs. *Philosophical transactions of the Royal Society of London Series B, containing papers of a biological character*, **213**, 21–87.

1.1. PHOTO ATTRIBUTION

Figures 1.1, 1.2, 1.3 1.14 and 1.15 are created by scripts that can be found at https://github.com/richelbilderbeek/thesis_introduction. Figure 1.4 is taken from https://commons.wikimedia.org/wiki/File:Darwin_Tree_1837.png. The drawing of Darwin's finches in figure 1.5 is taken from https://commons.wikimedia.org/wiki/File:Darwin%27s_finches_by_Gould.jpg. The evolution of Homoniniae in figure 1.6 is made by Dbachmann and taken from https://en.wikipedia.org/wiki/File:Hominini_lineage.svg. The phylogeny of figure 1.8 is by Aglondon, from https://commons.wikimedia.org/wiki/File:Edge_tree.png. The Largetooth Sawfish of figure 1.8 is taken from fromhttp://www.edgeofexistence.org/species/largetooth-sawfish The PHYLIP logo in figure 1.9 is taken from the PHYLIP homepage at http://evolution.genetics.washington.edu/phylip.html. The BEAST2 logo within figure 1.9, as well as the DensiTree picture are taken from the BEAST2 homepage at http://www.beast2.org. Figures 1.18, 1.20, 1.21, and 1.22 are actual screenshots from BEAUti v2.6.1. Figure 1.19 is from https://commons.wikimedia.org/wiki/File:Transitions_and_transversions.svg. The image of figure 1.23 is from https://commons.wikimedia.org/wiki/File:The_Earth_seen_from_Apollo_17.jpg.

2

BABETTE: BEAUTI 2, BEAST2 AND TRACER FOR R

Richèl J.C. Bilderbeek, Rampal S. Etienne

2

ABSTRACT

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.*

Samenvatting

1. *In de fylogenetica is BEAST2 een van de meest gebruikte hulpprogramma's. Het is gebundeld met de grafische gebruiksinterface BEAUti 2, DensiTree en Tracer, om BEAST2-configuratiebestanden te maken en om BEAST2-outputbestanden te interpreteren. Echter, als veel verschillende aligneringen of modelopzetten nodig zijn, is een werkvolgorde van meerdere grafische gebruiksinterfaces onhandig.*
2. *Hier presenteren we een gratis, vrij en open-source package,* `babette`*: 'BEAUti 2, BEAST2 en Tracer voor R', voor de programmeertaal R.* `babette` *schrijft BEAST2-configuratiebestanden, start BEAST2 and verwerkt de resultaten, alles met een enkele R functie-aanroep.*
3. *We beschrijven hoe* `babette` *te gebruiken is en de nieuwe mogelijkheden die het biedt vergeleken met de originele programma's, aan de hand van enkele voorbeelden.*
4. *Omdat* `babette` *ontworpen is voor uitbreidbaarheid en hoge kwaliteit, sluiten we af met het beschrijven van de verdere ontwikkeling van dit package.*

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

2.1. 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 can also estimate relevant parameters such as extinction and speciation rates from them. There are many phylogenetics tools available to obtain an estimate 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 phylogenies and model parameters, from one or more DNA, RNA or amino acid alignments (see figure 1 for an overview of the workflow).

BEAST2 has a graphical and a command-line interface, that both need a configuration 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 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 manageable 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), for a sliding-window analysis on a genomic alignment, or for a more thorough sensitivity analysis, one would like to loop through multiple (simulated or shortened) alignments, nucleotide substitution models, clock models and tree priors. One such tool to replace BEAUti is `BEASTmasteR` (Matzke 2015), 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 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 (BEAST1 and) 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. 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 to posterior analysis in R.

2.2. DESCRIPTION

babette is written in the R programming language (R Core Team 2013) and enables the full BEAST2 workflow from a single R function call, in a similar way to what subsequent usage of BEAUti, DensiTree and Tracer would produce. babette's main function is bbt_run, which configures BEAST2, runs it and parses its output. bbt_run needs at least the name of a FASTA file containing a DNA alignment. The default settings for the other arguments of bbt_run are identical to BEAUti's and BEAST2's default 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 the 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 contain all functionality in BEAUti, BEAST2 and their many plug-ins, because these tools themselves also change in time. babette currently works only on DNA data, because this is the most common use case. Nevertheless, babette provides the majority of default tree priors and supports the most important command-line arguments of BEAST2, provides the core Tracer analysis options, and has the most basic subset of plotting options of DensiTree. Up till now, the babette features implemented are those requested by users. Further extension of babette will be based on future user requests.

2.3. USAGE

babette can be installed easily from CRAN:

```
install.packages("babette")
```

For the most up-to-date version, one can download and install the package from babette's GitHub repository:

```
devtools::install_github("richelbilderbeek/babette")
```

To start using babette, load its functions in the global namespace first:

```
library(babette)
```

Because babette calls BEAST2, BEAST2 must be installed. This can be done from R, using:

```
install_beast2()
```

This will install BEAST2 to the default user data folder, but a different path can be specified as well. BEAUti, and likewise babette, needs at least a FASTA filename to produce a

BEAST2 configuration file. In BEAUti, this is achieved by loading a FASTA file, 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 storing the posterior. These output files must be parsed by either Tracer or DensiTree. In `babette`, all this is achieved by:

```
out <- bbt_run(fasta_filenames = "anthus_aco.fas")
```


This code will create a (temporary) BEAST2 configuration file, from the FASTA file with name `anthus_aco.fas` (which is supplied with the package, from Van Els & Norambuena 2018), using the same default settings as BEAUti, 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 estimates (called `estimates`), posterior phylogenies (called `anthus_aco_trees`, named after the alignment's name) and MCMC operator performance (`operators`). An example of using a different site model, clock model and tree prior is:

```
out <- bbt_run(
  fasta_filenames = "anthus_aco.fas",
  site_models = create_hky_site_model(),
  clock_models = create_rln_clock_model(),
  tree_priors = create_bd_tree_prior()
)
```

This code uses an HKY site model, a relaxed log-normal clock model and a birth-death tree prior, each with their default settings in BEAUti. Table 2.1 shows an overview of all functions to create site models, clock models and tree priors. Note 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 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, each with its own site model, is:

```
out <- bbt_run(
  fasta_filenames = c(
    "anthus_aco.fas",
    "anthus_nd2.fas"
  ),
  site_models = list(
    create_tn93_site_model(),
    create_gtr_site_model()
  )
)
```

`babette` also uses the same default prior distributions as BEAUti for each of the site models, clock models and tree priors. For example, by default, a Yule tree prior assumes that the birth rate follows a uniform distribution, from minus infinity to plus infinity. One may prefer a different distribution instead. Here is an example how to specify an exponential distribution for the birth rate in a Yule tree prior in `babette`:

```
out <- bbt_run(
  fasta_filenames = "anthus_aco.fas",
```

```
  tree_priors = create_yule_tree_prior(
    birth_rate_distr = create_exp_distr()
  )
)
```


In this same example, one may specify the initial shape parameters of the exponential distribution. In BEAST2's implementation, an exponential distribution has one shape parameter: its mean, which can be set to any value with `BEAUti`. To set the mean value of the exponential distribution to a fixed (non-estimated) value, do:

```
out <- bbt_run(
  fasta_filenames = "anthus_aco.fas",
  tree_priors = create_yule_tree_prior(
    birth_rate_distr = create_exp_distr(
      mean = create_mean_param(
        value = 1.0,
        estimate = FALSE
      )
    )
  )
)
```

`babette` also supports node dating. Like BEAUti, one can specify Most Recent Common Ancestor ('MRCA') priors. An MRCA prior allows to specify taxa having a common ancestor, including a distribution for the date of that ancestor. With `babette`, this is achieved as follows:

```
out <- bbt_run(
  fasta_filenames = "anthus_aco.fas",
  mrca_priors = create_mrca_prior(
    taxa_names = sample(get_taxa_names("anthus_aco.fas"), size
       = 2),
    alignment_id = get_alignment_id("anthus_aco.fas"),
    is_monophyletic = TRUE,
    mrca_distr = create_normal_distr(
      mean = create_mean_param(value = 15.0, estimate = FALSE),
      sigma = create_sigma_param(value = 0.025, estimate =
          FALSE)
    )
  )
)
```

Instead of dating the ancestor of two random taxa, any subset of taxa can be selected, and multiple sets are allowed. `babette` allows for the same core functionality as Tracer to show the values of the parameter estimates sampled in the BEAST2 run. This is called the "trace" (hence the name). The start of the trace, called the "burn-in", is usually discarded, as an MCMC algorithm (such as used by BEAST2) first has to converge to its equilibrium and hence the parameter estimates are not representative. By default, Tracer discards the first 10% of all the parameter estimates. To remove a 20% burn-in from all parameter estimates in `babette`, the following code can be used:

```
traces <- remove_burn_ins(
  traces = out$estimates,
  burn_in_fraction = 0.2
)
```


Tracer shows the ESSes of each posterior's variables. These ESSes are important 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 (of all estimated variables) in babette:

```
esses <- calc_esses(
  traces = traces,
  sample_interval = 1000
)
```

Tracer displays multiple summary statistics for each estimated variable: the mean and its standard error, standard deviation, variance, median, mode, geometric mean, 95% highest posterior density interval, auto-correlation time and effective sample size. It displays these statistics per variable. In babette, these summary statistics are collected for all estimated parameters at once:

```
sum_stats <- calc_summary_stats(
  traces = traces,
  sample_interval = 1000
)
```

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, allowing to see the uncertainty in topology and branch lengths. The posterior phylogenies are stored as anthus_aco_trees in the object out, and can be plotted as follows:

```
plot_densitree(phylos = out$anthus_aco_trees)
```

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 default, use:

```
create_beast2_input_file(
  input_filenames = babette::get_babette_path("anthus_aco.fas")
    ,
  output_filename = "beast2.xml"
)
```

This file can then be loaded and edited by BEAUti, run by BEAST2, or run by babette:

```
run_beast2(
  input_filename = "beast2.xml",
  output_log_filename = "run.log",
  output_trees_filenames = "posterior.trees",
  output_state_filename = "final.xml.state"
)
```

run_beast2 is a function that only runs BEAST2, and does not parse the output files (unlike bbt_run). In the example above, we specify the names of the desired BEAST2

output files explicitly, and these will be created in the R working directory, after which they can be inspected with other tools, or used to continue a BEAST2 run. When the names of these files are not specified, both `bbt_run` and `run_beast2` put these files in the default temporary folder (as obtained from `temp.dir()`) to keep the working directory clean of intermediate files.

2.4. BABETTE RESOURCES

`babette` is free, libre and open source software available at

`http://github.com/richelbilderbeek/babette`

and is licensed under the GNU General Public License v3.0. `babette` uses the Travis CI (`https://travis-ci.org`) continuous integration service, which is known to significantly increase the number of bugs exposed (Vasilescu *et al.* 2015) and increases the speed at which new features are added (Vasilescu *et al.* 2015). `babette` has a 100% code coverage, which correlates with code quality (Del Frate *et al.* 1995, Horgan *et al.* 1994). `babette` follows Hadley Wickham's style guide (Wickham 2015), which improves software quality (Fang 2001). `babette` depends on multiple packages, which are `ape` (Paradis *et al.* 2004), `beautier` (Bilderbeek 2018b), `beastier` (Bilderbeek 2018a), `devtools` (Wickham & Chang 2016), `geiger` (Harmon *et al.* 2008), `ggplot2` (Wickham 2009), `knitr` (Xie 2017), `phangorn` (Schliep 2011), `rmarkdown` (Allaire *et al.* 2017), `seqinr` (Charif & Lobry 2007), `stringr` (Wickham 2017), `testit` (Xie 2014) and `tracerer` (Bilderbeek 2018c). We tested `babette` to give a clean error message for incorrect input, by calling `babette` one million times with random or random sensible inputs, using a high performance computer cluster. The test scripts 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 improves transparency (Gorgolewski & Poldrack 2016). `babette`'s GitHub facilitates feature requests and has guidelines how to do so.

`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 documentation links to related functions. Additionally, `babette` has a vignette that demonstrates extensively how to use it. There is documentation on the GitHub to get started, with a dozen examples of BEAUti screenshots with equivalent `babette` code. Finally, `babette` has tutorial videos that can be downloaded or viewed on YouTube, `https://goo.gl/weKaaU`.

2.5. CITATION OF BABETTE

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")
```

2.6. 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, Michael Matschiner and one anonymous reviewer for reviewing this article. Thanks to rOpenSci, and especially Noam Ross and Guangchuang Yu 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 cluster. We thank the Netherlands Organization for Scientific Research (NWO) for financial support through a VICI grant awarded to RSE.

2.7. DATA ACCESSIBILITY

All code is archived at `http://github.com/richelbilderbeek/babette_article`, with DOI `https://doi.org/10.5281/zenodo.1251203`.

2.8. 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 substantially to revisions.

REFERENCES

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J. & Chang, W. (2017) *rmarkdown: Dynamic Documents for R*. R package version 1.8.

Bilderbeek, R.J. (2018a) . `https://github.com/richelbilderbeek/beastier` [Accessed: 2018-03-16].

Bilderbeek, R.J. (2018b) . `https://github.com/richelbilderbeek/beautier` [Accessed: 2018-03-16].

Bilderbeek, R.J. (2018c) . `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.

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., *Structural approaches to sequence evolution: Molecules, networks, populations*, Biological and Medical Physics, Biomedical Engineering, pp. 207–232. Springer 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 Engineering, 1995. Proceedings., Sixth International Symposium on*, pp. 124–132. IEEE.

Drummond, A.J., Suchard, M.A., Xie, D. & Rambaut, A. (2012) Bayesian phylogenetics 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. IEEE.

Faria, N. & Suchard, M.A. (2015) . `https://github.com/beast-dev/RBeast` [Accessed: 2018-03-02].

Gorgolewski, K.J. & Poldrack, R. (2016) A practical guide for improving transparency 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., Leprevost, 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) . `https://github.com/olli0601/rBEAST` [Accessed: 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) . *Bioinformatics*, **27**, 592–593.

Van Els, P. & Norambuena, H.V. (2018) A revision of species limits in neotropical pipits anthus based on multilocus genetic and vocal data. *Ibis*.

Vasilescu, B., Yu, Y., Wang, H., Devanbu, P. & Filkov, V. (2015) Quality and productivity outcomes relating to continuous integration in github. *Proceedings 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 Packages Easier*. R package version 1.12.0.9000.

Xie, Y. (2014) *testit: A Simple Package for Testing R Packages*. R package version 0.4, `http://CRAN.R-project.org/package=testit`.

Xie, Y. (2017) *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.17.

2

Name	Description
bbt_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 2.1 | babette's main functions

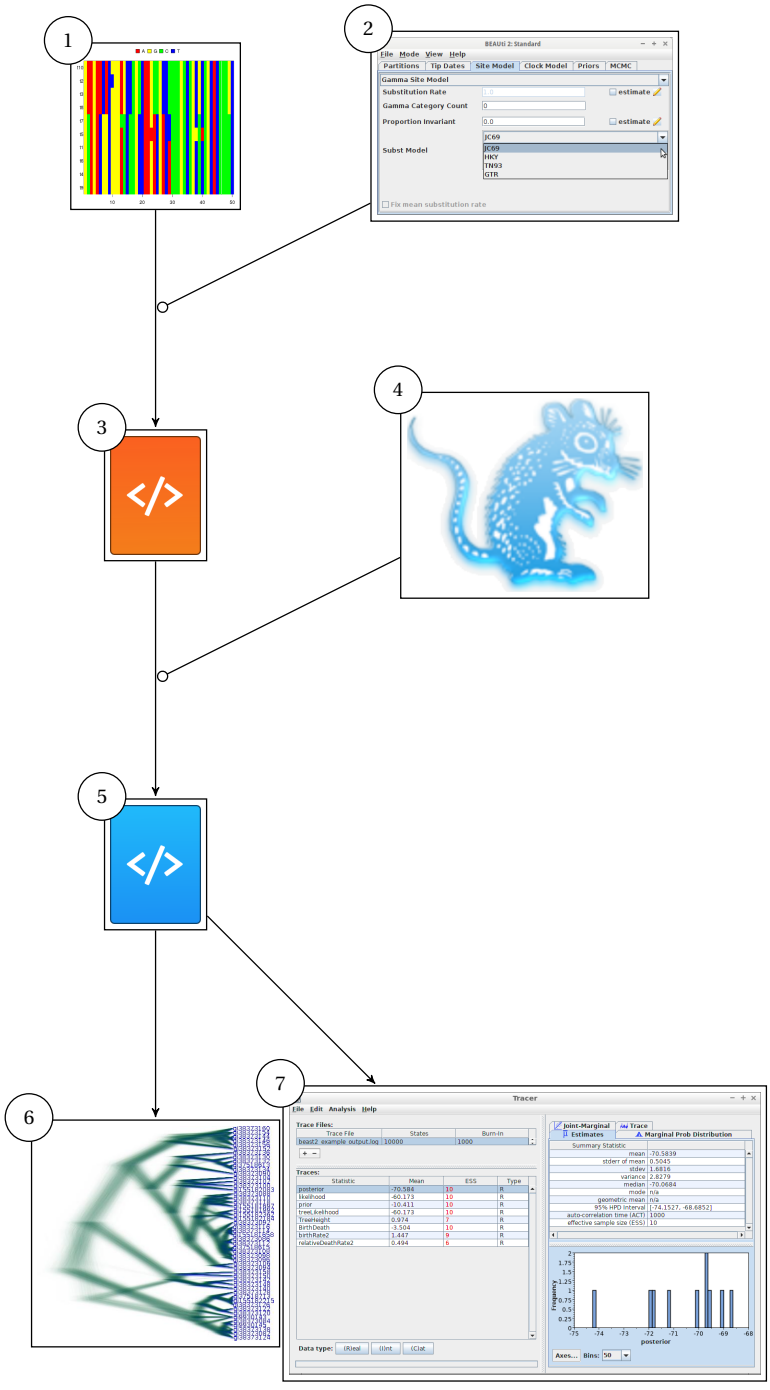


Figure 2.1 | Workflow using GUI tools. From an alignment (1) and BEAUti (2), a BEAST2 configuration file (3) is created. BEAST2 (4) uses that file to infer a posterior, storing it in multiple files (5). These results are visualized using DensiTree (6) and Tracer (7). `babette` allows for the same workflow, all from an R function call

3

PIROUETTE

Richèl J.C. Bilderbeek, Giovanni Laudanno, Rampal S. Etienne

ABSTRACT

1. *Phylogenetic trees are currently routinely reconstructed from an alignment of character sequences (usually nucleotide sequences). Bayesian tools, such as MrBayes, RevBayes and BEAST2, have gained much popularity over the last decade, as they allow joint estimation of the posterior distribution of the phylogenetic trees and the parameters of the underlying inference model. An important ingredient of these Bayesian approaches is the species tree prior. In principle, the Bayesian framework allows for comparing different tree priors, which may elucidate the macroevolutionary processes underlying the species tree. In practice, however, only macroevolutionary models that allow for fast computation of the prior probability are used. The question is how accurate the tree estimation is when the real macroevolutionary processes are substantially different from those assumed in the tree prior.*

2. *Here we present* `pirouette`*, a free and open-source R package that assesses the inference error made by Bayesian phylogenetics for a given macroevolutionary diversification model.* `pirouette` *makes use of BEAST2, but its philosophy applies to any Bayesian phylogenetic inference tool.*

3. *We describe* `pirouette`*'s usage providing full examples in which we interrogate a model for its power to describe another.*

4. *Last, we discuss the results obtained by the examples and their interpretation.*

Keywords: Bayesian model selection, BEAST2, computational biology, evolution, phylogenetics, R, tree prior, babette

3.1. INTRODUCTION

The development of new powerful Bayesian phylogenetic inference tools, such as BEAST [Drummond & Rambaut 2007], MrBayes [Huelsenbeck & Ronquist 2001] or RevBayes [Höhna *et al.* 2016] has been a major advance in constructing phylogenetic trees from character data (usually nucleotide sequences) extracted from organisms (usually extant, but extinction events and/or time-stamped data can also be added), and hence in our understanding of the main drivers and modes of diversification.

BEAST [Drummond & Rambaut 2007] is a typical Bayesian phylogenetics tool, that needs both character data and priors to infer a posterior distribution of phylogenies. Specifically, for the species tree prior - which describes the process of diversification - BEAST has built-in priors such as the Yule [Yule 1925] and (constant-rate) birth-death (BD) [Nee *et al.* 1994] models as well as coalescent priors. These simple tree priors are among the most commonly used, as they represent some biologically realistic processes (e.g. viewing diversification as a branching process), while being computationally fast.

To allow users to extend the functionalities of BEAST using plug-ins, BEAST2 was written [Bouckaert *et al.* 2019] (with BEAST and BEAST2 still independently being developed further). For example, one can add novel diversification models by writing a BEAST2 plugin that contains the likelihood formula of a phylogeny under the novel diversification model, i.e. the prior probability of a species tree. Plugins have been provided, for instance, for the calibrated Yule model [Heled & Drummond 2015], the BD model with incomplete sampling [Stadler 2009], the BD model with serial sampling [Stadler *et al.* 2012], the BD serial skyline model [Stadler *et al.* 2013], the fossilized BD process [Gavryushkina *et al.* 2014], and the BD SIR model [Kühnert *et al.* 2014].

Many other diversification models (and their associated likelihood algorithms) have been developed, e.g., models in which diversification is time-dependent [Nee *et al.* 1994, Rabosky & Lovette 2008], or diversity-dependent [Etienne *et al.* 2012], or where diversification rates change for specific lineages and their descendants [Alfaro *et al.* 2009, Etienne & Haegeman 2012, Laudanno *et al.* 2020, Rabosky 2014]. Other models treat speciation as a process that takes time [Etienne & Rosindell 2012, Lambert *et al.* 2015, Rosindell *et al.* 2010], or where diversification rates depends on one or more traits [FitzJohn 2012, Herrera-Alsina *et al.* 2019, Maddison *et al.* 2007].

These are, however, not yet available as tree priors in BEAST2, for reasons explained below. In this paper, we present methodology to determine whether such new plug-ins are needed, or whether currently available plug-ins are sufficient. We show this using the Yule and BD species tree priors, but our methods can be used with other built-in tree priors as well.

The rationale of our paper is as follows. When a novel diversification model is introduced, its performance in inference should be tested. Part of a model's performance is its ability to recover parameters from simulated data with known parameters (e.g. [Etienne *et al.* 2014]), where ideally the estimated parameter values closely match the known/true values. Even when a diversification model passes this test, it is not necessarily used as tree prior in Bayesian inference. Bayesian phylogenetic inference often requires that the prior probability of the phylogeny according to the diversification model has to be computed millions of times. Therefore, biologically interesting but computationally expensive tree priors are often not implemented, and simpler priors are used instead. This is not necessarily problematic, when the data are very informative or when the prior is truly uninformative, as this will reduce the influence of the tree prior. However, the assumption that tree prior choice is of low impact must first be verified.

There have been multiple attempts to investigate the impact of tree prior choice. For example, Sarver and colleagues, [Sarver *et al.* 2019] showed that the choice of tree prior does not substantially affect phylogenetic inferences of diversification rates. However, they only compared current diversification models to one another, and thus this does not inform us on the impact of a new tree prior.

Similarly, Ritchie and colleagues [Ritchie *et al.* 2016] showed that inference was accurate when birth-death or skyline coalescent priors were used, but they simulated their trees with a Yule process only, as their focus was not so much on the diversification process but on the influence of inter- and intraspecific sampling.

Another way to benchmark a diversification model, is by doing a model comparison, in which the best model is determined from a set of models. A good early example is

Goldman 1993 in which Goldman compared DNA substitution models. A recent approach to test the impact of tree prior choice, proposed by Duchene *et al.* 2018, allows to measure model adequacy for phylodynamic models that are mathematically described (i.e. have a known likelihood equation).

Here we introduce a method to quantify the impact of a novel tree prior, i.e., a tree model, for which we can simulate phylogenies, but not yet calculate their likelihoods. This new method simultaneously assesses the substitution, clock and tree models [Duchêne *et al.* 2015]. The method starts with a phylogeny generated by the new model. Next, nucleotide sequences are simulated that follow the evolutionary history of the given phylogeny. Then, using BEAST2's built-in tree priors, a Bayesian posterior distribution of phylogenies is inferred. We then compare the inferred with the original phylogenies. How to properly perform this comparison forms the heart of our method. Only new diversification models that result in a large discrepancy between inferred and simulated phylogenies will be worth the effort and computational burden to implement a species tree prior for in a Bayesian framework.

Our method is programmed as an R package [R Core Team 2013] called `pirouette`. `pirouette` is built on `babette` [Bilderbeek & Etienne 2018], which calls BEAST2 [Bouckaert *et al.* 2019].

3.2. DESCRIPTION

The goal of `pirouette` is to quantify the impact of a new tree prior. It does so by measuring the inference error made for a given reconstructed phylogeny, simulated under a (usually novel) diversification model. We refer to the model that has generated the given tree as the 'generative tree model' p_G. A 'generative tree model', in this paper, can be either the novel diversification model for which we are testing the impact of choosing standard tree priors for, or it is the model with which we generate the twin tree that is needed for comparison (see below). In the latter case, we also refer to it as the actual generative tree model, and it thus serves as a baseline model. This is is done in the example, where the Yule model is the generative model.

The inference error we aim to quantify is not of stochastic nature. Stochastic errors are usually non-directional. We, instead, aim to expose the bias due to the mismatch between a generative model (that has generated the phylogeny) and the model(s) used in the actual inference. We define the birth-death (BD) model [Nee *et al.* 1994] as the standard tree model, as many (non-standard) tree models have a parameter setting such that it reduces to this model. One such example is the diversity-dependent (DD) diversification model [Etienne & Haegeman 2020, Etienne *et al.* 2012] in which speciation or extinction rate depends on the number of species and a clade-level carrying capacity. The BD model can be seen as a special case of the DD model, because for an infinite carrying capacity, the DD model reduces to the BD model. When benchmarking a novel tree model, one will typically construct phylogenies for different combinations of the diversification model's parameters, to assess under which scenarios the inference error cannot be neglected. While we recommend many replicate simulations when assessing a novel tree prior, our example contains only one replicate, as the goal is to show the workings of `pirouette`, instead of doing an extensive analysis. The supplementary material includes results of

replicated runs under multiple settings.

`pirouette` allows the user to specify a wide variety of custom settings. These settings can be grouped in macro-sections, according to how they operate in the pipeline. We summarize them in Table 3.1 and Table 3.2.

3.2.1. PIROUETTE'S PIPELINE

The pipeline to assess the error BEAST2 makes in inferring this phylogeny contains the following steps:

1. The user supplies one or (ideally) more phylogenies from a new diversification model.
2. From the given phylogeny an alignment is simulated under a known alignment model A.
3. From this alignment, according to the specified inference conditions C, an inference model I is chosen (which may or may not differ from the model that generated the tree).
4. The inference model and the alignment are used to infer a posterior distribution of phylogenies.
5. The phylogenies in the posterior are compared with the given phylogeny to estimate the error made, according to the error measure E specified by the user.

The pipeline is visualized in Fig. 3.1. There is also the option to generate a 'twin tree', that goes through the same pipeline (see supplementary subsection 3.5.5).

The first step simulates an alignment from the given phylogeny (Fig. 3.1, 1a → 2a). For the sake of clarity, here we will assume the alignment consists of DNA sequences, but one can also use other heritable materials such as amino acids. The user must specify a root sequence (i.e. the DNA sequence of the shared common ancestor of all species), a mutation rate and a site model.

The second step (Fig. 3.1, 3a) selects one or more inference model(s) I from a set of standard inference models $I_1, \ldots, I_n$. For example, if the generative model is known and standard (which it is for the twin tree, see below), one can specify the inference model to be the same as the generative model. If the tree model is unknown or non-standard - which is the primary motivation for this paper -, one can pick a standard inference model which is considered to be closest to the true tree model. Alternatively, if we want to run only the inference model that fits best to an alignment from a set of candidates (regardless of whether these generated the alignments), one can specify these inference models (see section 3.5.6).

The third step infers the posterior distributions, using the simulated alignment (Fig. 3.1, 2a → 4a), and the inference models that were selected in the previous step (3a). For each selected experiment a posterior distribution is inferred, using the `babette` [Bilderbeek & Etienne 2018] R package which makes use of BEAST2.

The fourth step quantifies the newimpact of choosing standard models for inference, i.e. the inference error made. First the burn-in fraction is removed, i.e. the first phase

Sub-argument	Description	Possible values
`tree_prior`	Macroevolutionary diversification model	BD, CBS, CCP, CEP, Yule
`clock_model`	Clock for the DNA mutation rates	RLN, strict
`site_model`	Nucleotide substitution model	GTR, HKY, JC, TN
`mutation_rate`	Pace at which substitutions occur	`mutation_rate` $\in \mathbb{R}_{>0}$
`root_sequence`	DNA sequence at the root of the tree	any combination of a, c, g, t
`model_type`	Criterion to select an inference model	Generative, Candidate
`run_if`	Condition under which an inference model is used	Always, Best candidate
`do_measure_evidence`	Sets whether or not the evidence of the model must be computed	TRUE, FALSE
`error_fun`	Specifies how to measure the error	nLTT, $\lvert\gamma\rvert$
`burn_in_fraction`	Specifies the percentage of initial posterior trees to discard	`burn_in_fraction` $\in [0, 1]$

Table 3.1 | Most important parameter options. BD = birth death [Nee *et al.* 1994], CBS = coalescent Bayesian skyline [Drummond *et al.* 2005], CCP = coalescent constant-population, CEP = coalescent exponential-population, Yule = pure birth model [Yule 1925], RLN = relaxed log-normal clock model [Drummond *et al.* 2006], strict = strict clock model [Zuckerkandl & Pauling 1965], GTR = Generalized time-reversible model [Tavaré 1986], HKY = Hasegawa, Kishino and Yano [Hasegawa *et al.* 1985], JC = Jukes and Cantor [Jukes *et al.* 1969], TN = Tamura and Nei [Tamura & Nei 1993], nLTT = normalized lineages-through-time [Janzen *et al.* 2015], $|\gamma|$ = absolute value of the gamma statistic [Pybus & Harvey 2000].

Symbol	Macro-argument	Description
G	Generative model	The full setting to produce BEAST2 input data. Its core features are the tree prior p_G, the clock model c_G and the site model s_G.
s_G	Site model	Both the substitution model and rate variation across sites.
A	Alignment model	Specifies the alignment generation, such as the clock model c_G, site model s_G and root sequence.
X_i	i-th candidate experiment	Full setting for a Bayesian inference. It is made by a candidate inference model I_i and its inference conditions C_i.
I	Inference model	The assumed phylogenetic inference model, of which the main components are the tree prior p_I, assumed clock model c_I and assumed site model s_I.
C	Inference conditions	Conditions under which I is used in the inference. They are composed of the model type, run condition and whether to measure the evidence.
E	Error measure parameters	Errors measurement setup that can be specified providing an error function to measure the difference between the original phylogeny and the inferred posterior. The first iterations of the MCMC chain of the posterior may not be representative and can be discarded using a burn-in fraction.

Table 3.2 | Definitions of terms and relative symbols used in the main text and in Fig 3.1. To run the pipeline A, X and E must be specified.

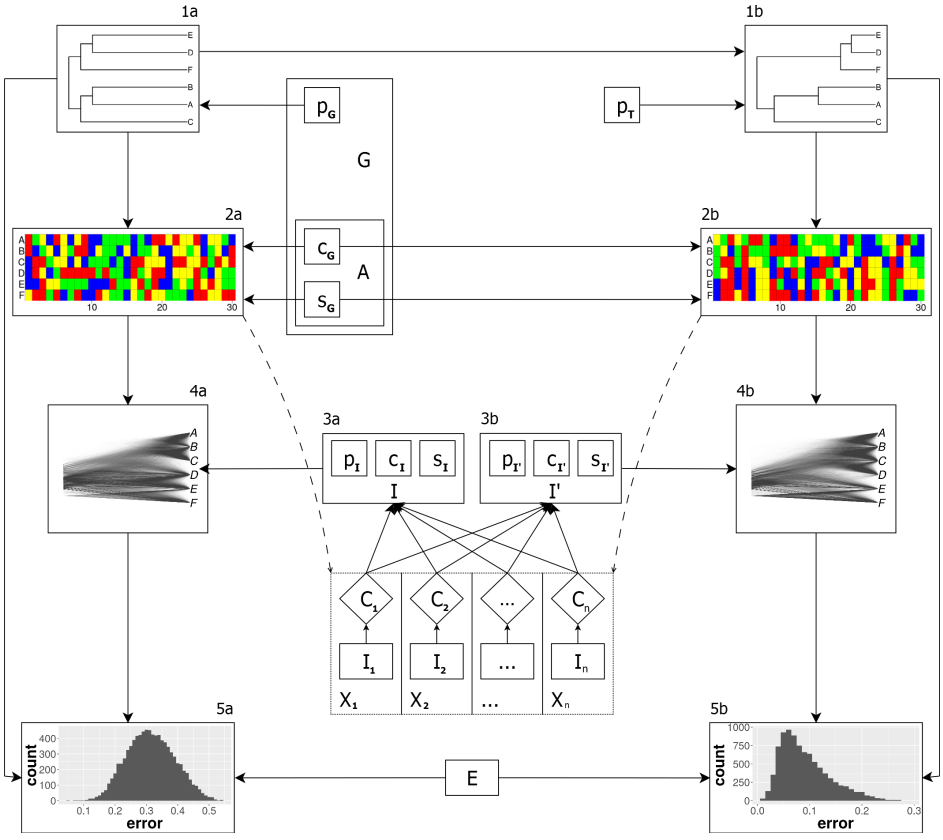


Figure 3.1 | `pirouette` pipeline. The pipeline starts from a phylogeny (1a) simulated by the generative tree model p_G. The phylogeny is converted to an alignment (2a) using the generative alignment model $A = (c_G, s_G)$, composed of a clock model and a site model. The user defines one or more experiments. For each candidate experiment X_i (a combination of inference model I_i and condition C_i), if its condition C_i is satisfied (which can depend on the alignment), the corresponding inference model $I = I_i$ is selected to be used in the next step. The inference models (3a) of the selected experiments use the alignment (2a) to each create a Bayesian posterior of (parameter estimates and) phylogenies (4a). Each of the posterior trees is compared to the true phylogeny (1a) using the error measure E, resulting in an error distribution (5a). Optionally, for each selected inference model a twin pipeline can be run. A twin phylogeny (1b) can be generated from the original phylogeny (1a) using the twin tree model p_t, selected among standard diversification models; the default option is the standard BD model, with parameters estimated from the original phylogeny. A twin alignment (2b) is then simulated from the twin phylogeny using clock model c_G and site model s_G used with the generative tree model (the novel tree model). The twin pipeline follows the procedure of the main pipeline, resulting in a twin error distribution (5b).

of the Markov chain Monte Carlo (MCMC) run, which samples an unrepresentative part of parameter and tree space. From the remaining posterior, `pirouette` creates an error distribution, by measuring the difference between the true tree and each of the posterior trees (Fig. 3.1, 4a → 5a). The user can specify a function to quantify the differences between the true and posterior trees.

3.2.2. CONTROLS

`pirouette` allows for two types of control measurements. The first type of control is called 'twinning', which results in an error distribution that is the baseline error of the inference pipeline (see supplementary materials, subsection 3.5.5 for more details). This the error that arises when the models used in inference are identical to the ones used in generating the alignments. The second type of control is the use of candidate models, which result in an error distribution for a generative model that is determined to be the best fit to the tree (see supplementary materials, section 3.5.6 for more details). The underlying idea is that using a substitution model in inference than used in generating the alignment may partly compensate for choosing a standard tree model instead of the generative tree model as tree prior in inference. Additionally, multiple `pirouette` runs are needed to reduce the influence of stochasticity (see supplementary materials, section 3.5.7 for more details).

3.3. USAGE

We show the usage of `pirouette` on a tree generated by the non-standard diversity-dependent (DD) tree model [Etienne & Haegeman, 2020, Etienne *et al.*, 2012], which is a BD model with a speciation rate that depends on the number of species.

The code to reproduce our results can be found at `https://github.com/richelbilderbeek/pirouette_example_30` and a simplified version is shown here for convenience:

```
library(pirouette)

# Create a DD phylogeny with 5 taxa and a crown age of 10
phylogeny <- create_exemplary_dd_tree()

# Use standard pirouette setup. This creates a list object with
    all settings for generating the alignment, the inference
   using BEAST2, the twinning parameters to generate the twin
   tree and infer it using BEAST2, and the error measure
pir_params <- create_std_pir_params()

# Do the runs
pir_out <- pir_run(
  phylogeny = phylogeny,
  pir_params = pir_params
)

# Plot
```

```
pir_plot(pir_out)
```

The DD tree generated by this code is shown in Figure 3.2.

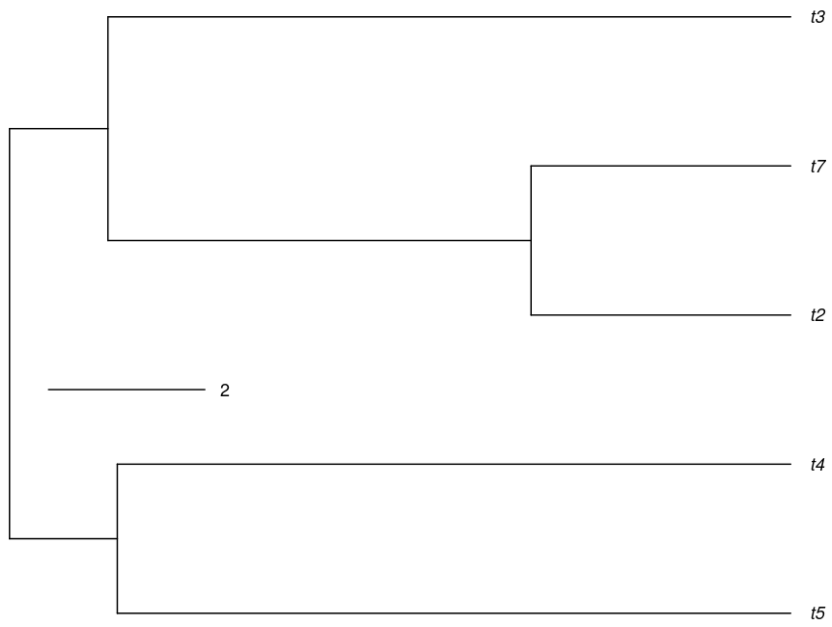


Figure 3.2 | The example tree resulting from a diversity-dependent (DD) simulation.

The error distribution shown in Figure 3.3 is produced, which uses the nLTT statistic [Janzen *et al.* 2015] to compare phylogenies (see section 3.5.8 for details regarding the nLTT statistic and its caveats).

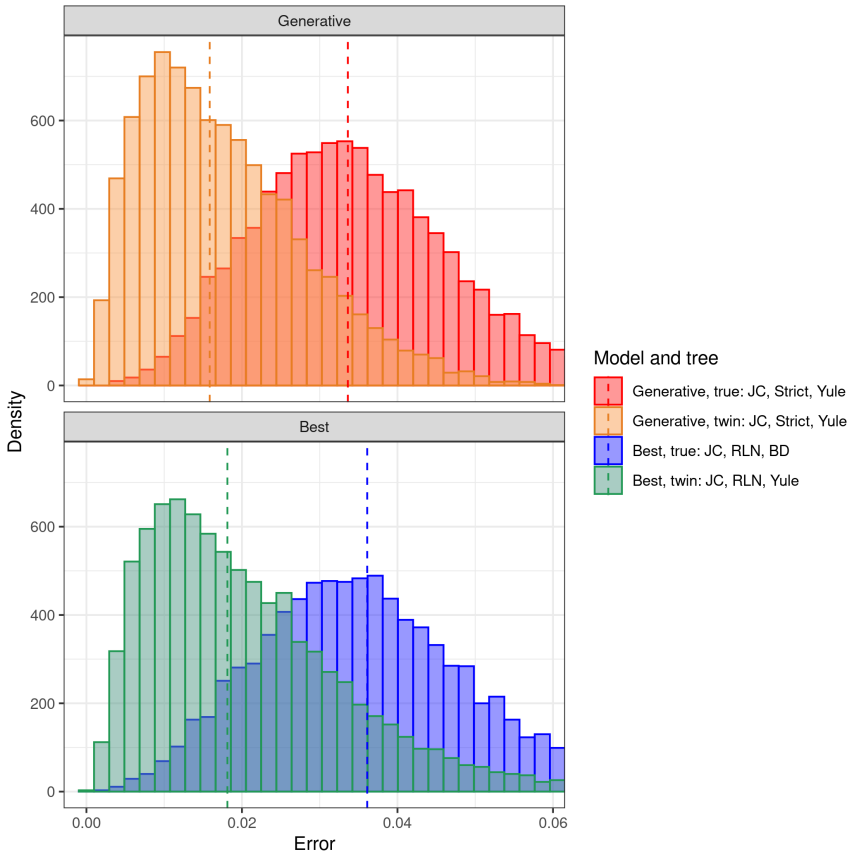


Figure 3.3 | The impact of the tree prior for the example tree in Figure 3.2. The alignment for this true tree was generated using a JC substitution model and strict clock model. For inferring the tree from this alignment in the 'generative' scenario, the same substitution and clock models were used, and a Yule tree prior (this is the assumed generative model, because the real generative model is assumed to be unknown). For the twin tree, the same inference models were used. In the 'best' scenario, for the true tree, the best-fitting candidate models were JC substitution model, RLN clock model and BD tree prior, while for the twin tree, the best-fitting candidate models were JC substitution model, RLN clock model and Yule tree prior. The twin distributions show the baseline inference error. Vertical dashed lines show the median error value per distribution.

In the upper panel of Figure 3.3, we can see that the error distributions of the (assumed) generative model (i.e. the known generative substitution and clock models, and the tree model that is assumed in inference of the true tree, and the tree model that is used for generating and inferring the twin tree) differ substantially between the true and twin tree. This difference shows the extent of the mismatch between the true tree model (which is DD) and the (Yule) tree prior used in inference. Because these distributions are distinctively different, the inference error made when using an incorrect tree prior on a DD tree is quite profound.

Comparing the upper and lower panel of Figure 3.3, we can see that the best candidate model is slightly worse at inferring the true tree, than the (assumed) generative model,

indicating that the generative inference model we selected is a good choice.

The candidate model that had highest evidence given the simulated alignment, was JC, RLN and BD (see Table 3.1 for the meaning of these abbreviations). The RLN clock model is a surprising result: it assumes nucleotide substitutions occur at different rates between the taxa. The JC nucleotide substitution model matches the model used to simulate the alignment. The BD model is perhaps somewhat surprising for the true tree, because the other alternative standard tree prior, Yule, is probably closest to the true DD model because it shows no pull-of-the-present (but also no slowdown).

3.4. DISCUSSION

We showed how to use `pirouette` to quantify the impact of a tree prior in Bayesian phylogenetics, assuming - for illustrative purposes - the simplest standard substitution, clock and tree models, but also the models that would be selected among many different standard tree priors according to the highest marginal likelihood, as this would be a likely strategy for an empiricist. We recommend exploring different candidate models, but note that this is computationally highly demanding, particularly for large trees.

Figure 3.3 illustrates the primary result of our pipeline: it shows the error distributions for the true tree and the twin tree when either the generative model (for substitution and clock models these are known, for the tree model it must be assumed for the true tree and it is known for the twin tree) or the best-fitting set candidate model (i.e. combination of tree model, substitution model and clock model) is used in inference. The clear difference between the error distributions for the true tree and the twin tree suggests that the choice of tree prior matters. We note, however, that only one tree from a novel tree model is not enough to determine the impact of using an incorrect tree prior. Instead, a distribution of multiple trees, generated by the novel tree model, should be used. In the supplementary material we have provided some examples.

Like most phylogenetic experiments, the setup of `pirouette` involves many choices. A prime example is the length of the simulated DNA sequence. One expects that the inference error decreases for longer DNA sequences. We investigated this superficially and confirmed this prediction (see the supplementary material). However, we note that for longer DNA sequences, the assumption of the same substitution rates across the entire sequence may become less realistic (different genes may experience different substitution rates) and hence longer sequences may require more parameters. Hence, simply getting longer sequences will not always lead to a drastic reduction of the influence of the species tree prior. Fortunately, `pirouette` provides a pipeline that works for all choices.

Interpreting the results of `pirouette` is up to the user; `pirouette` does not answer the question whether the inference error is too large to trust the inferred tree. The user is encouraged to use different statistics to measure the error. The nLTT statistic is a promising starting point, as it can compare any two trees and results in an error distribution of known range, but one may also explore other statistics, for example statistics that depend on the topology of the tree, While `pirouette` allows for this in principle, in our example we used a diversification model (DD) that only deviates from the Yule and BD models in the temporal branching pattern, not in the topology. For models that make different predictions on topology, the twinning process should be modified.

As noted in the introduction, Duchêne and colleagues [Duchene *et al.* 2018] also developed a method to assess the adequacy of a tree model on empirical trees. They simulated trees from the posterior distribution of the parameters and then compared this to the originally inferred tree using tree statistics, to determine whether the assumed tree model in inference indeed generates the tree as inferred. This is useful if these trees match, but when they do not, this does not mean that the inferred tree is incorrect; if sufficient data is available the species tree prior may not be important, and hence inference may be adequate even though the assumed species tree prior is not. In short, the approach is applied to empirical trees and compares the posterior and prior distribution of trees (with the latter generated with the posterior parameters!). By contrast, `pirouette` aims to expose when assuming standard priors for the species tree are a mis- or underparameterization. Hence, our approach applies to simulated trees and compares the posterior distributions of trees generated with a standard and non-standard model, but inferred with a standard one. The two methods therefore complement one another.

Furthermore, we note that the `pirouette` pipeline is not restricted to exploring the effects of a new species tree model. The pipeline can also be used to explore the effects of non-standard clock or site models, such as relaxed clock models with a non-standard distribution, correlated substitutions on sister lineages, or elevated substitutions rates during speciation events. It is, however, beyond the scope of this paper to discuss all these options in more detail.

In conclusion, `pirouette` can show the errors in phylogenetic reconstruction expected when the model assumed in inference is different from the actual generative model. The user can then judge whether or not this new model should be implemented in a Bayesian phylogenetic tool.

REFERENCES

Alfaro, M.E., Santini, F., Brock, C., Alamillo, H., Dornburg, A., Rabosky, D.L., Carnevale, G. & Harmon, L.J. (2009) Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proceedings of the National Academy of Sciences of the USA*, **106**, 13410–13414.

Bilderbeek, R.J. & Etienne, R.S. (2018) 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. *et al.* (2019) BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, **15**, e1006650.

Drummond, A.J., Ho, S.Y., Phillips, M.J. & Rambaut, A. (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biology*, **4**, e88.

Drummond, A.J. & Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, **7**, 214.

Drummond, A.J., Rambaut, A., Shapiro, B. & Pybus, O.G. (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, **22**, 1185–1192.

Duchêne, D.A., Duchêne, S., Holmes, E.C. & Ho, S.Y. (2015) Evaluating the adequacy of molecular clock models using posterior predictive simulations. *Molecular Biology and Evolution*, **32**, 2986–2995.

Duchene, S., Bouckaert, R., Duchene, D.A., Stadler, T. & Drummond, A.J. (2018) Phylodynamic model adequacy using posterior predictive simulations. *Systematic Biology*, **68**, 358–364.

Etienne, R.S. & Haegeman, B. (2012) A conceptual and statistical framework for adaptive radiations with a key role for diversity dependence. *The American Naturalist*, **180**, E75–E89.

Etienne, R.S. & Haegeman, B. (2020) . `https://CRAN.R-project.org/package=DDD`.

Etienne, R.S., Haegeman, B., Stadler, T., Aze, T., Pearson, P.N., Purvis, A. & Phillimore, A.B. (2012) Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 1300–1309.

Etienne, R.S., Morlon, H. & Lambert, A. (2014) Estimating the duration of speciation from phylogenies. *Evolution*, **68**, 2430–2440.

Etienne, R.S. & Rosindell, J. (2012) Prolonging the past counteracts the pull of the present: protracted speciation can explain observed slowdowns in diversification. *Systematic Biology*, **61**, 204–213.

FitzJohn, R.G. (2012) Diversitree: comparative phylogenetic analyses of diversification in R. *Methods in Ecology and Evolution*, **3**, 1084–1092.

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

Goldman, N. (1993) Statistical tests of models of dna substitution. *Journal of molecular evolution*, **36**, 182–198.

Hasegawa, M., Kishino, H. & Yano, T.a. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of Molecular Evolution*, **22**, 160–174.

Heled, J. & Drummond, A.J. (2015) Calibrated birth–death phylogenetic time-tree priors for bayesian inference. *Systematic Biology*, **64**, 369–383.

Herrera-Alsina, L., van Els, P. & Etienne, R.S. (2019) Detecting the dependence of diversification on multiple traits from phylogenetic trees and trait data. *Systematic Biology*, **68**, 317–328.

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

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

Janzen, T., Höhna, S. & Etienne, R.S. (2015) Approximate Bayesian computation of diversification rates from molecular phylogenies: introducing a new efficient summary statistic, the nLTT. *Methods in Ecology and Evolution*, **6**, 566–575.

Jukes, T.H., Cantor, C.R. *et al.* (1969) Evolution of protein molecules. *Mammalian Protein Metabolism*, **3**, 132.

Kühnert, D., Stadler, T., Vaughan, T.G. & Drummond, A.J. (2014) Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth–death sir model. *Journal of the Royal Society Interface*, **11**, 20131106.

Lambert, A., Morlon, H. & Etienne, R.S. (2015) The reconstructed tree in the lineage-based model of protracted speciation. *Journal of Mathematical Biology*, **70**, 367–397.

Laudanno, G., Haegeman, B., Rabosky, D.L. & Etienne, R.S. (2020) Detecting lineage-specific shifts in diversification: A proper likelihood approach. *Systematic Biology*.

Maddison, W.P., Midford, P.E. & Otto, S.P. (2007) Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, **56**, 701–710.

Nee, S., May, R.M. & Harvey, P.H. (1994) The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London B*, **344**, 305–311.

Pybus, O.G. & Harvey, P.H. (2000) Testing macro–evolutionary models using incomplete molecular phylogenies. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **267**, 2267–2272.

R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabosky, D.L. (2014) Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLOS One*, **9**, e89543.

Rabosky, D.L. & Lovette, I.J. (2008) Explosive evolutionary radiations: decreasing speciation or increasing extinction through time? *Evolution: International Journal of Organic Evolution*, **62**, 1866–1875.

Ritchie, A.M., Lo, N. & Ho, S.Y.W. (2016) The Impact of the Tree Prior on Molecular Dating of Data Sets Containing a Mixture of Inter- and Intraspecies Sampling. *Systematic Biology*, **66**, 413–425.

Rosindell, J., Cornell, S.J., Hubbell, S.P. & Etienne, R.S. (2010) Protracted speciation revitalizes the neutral theory of biodiversity. *Ecology Letters*, **13**, 716–727.

Sarver, B.A., Pennell, M.W., Brown, J.W., Keeble, S., Hardwick, K.M., Sullivan, J. & Harmon, L.J. (2019) The choice of tree prior and molecular clock does not substantially affect phylogenetic inferences of diversification rates. *PeerJ*, **7**, e6334.

Stadler, T. (2009) On incomplete sampling under birth–death models and connections to the sampling-based coalescent. *Journal of theoretical biology*, **261**, 58–66.

Stadler, T., Kouyos, R., von Wyl, V., Yerly, S., Böni, J., Bürgisser, P., Klimkait, T., Joos, B., Rieder, P., Xie, D. *et al.* (2012) Estimating the basic reproductive number from viral sequence data. *Molecular biology and evolution*, **29**, 347–357.

Stadler, T., Kühnert, D., Bonhoeffer, S. & Drummond, A.J. (2013) Birth–death skyline plot reveals temporal changes of epidemic spread in hiv and hepatitis c virus (hcv). *Proceedings of the National Academy of Sciences*, **110**, 228–233.

Tamura, K. & Nei, M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees. *Molecular biology and Evolution*, **10**, 512–526.

Tavaré, S. (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, **17**, 57–86.

Yule, G.U. (1925) A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *Philosophical transactions of the Royal Society of London Series B, containing papers of a biological character*, **213**, 21–87.

Zuckerkandl, E. & Pauling, L. (1965) Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, **8**, 357–366.

3.5. SUPPLEMENTARY MATERIAL

This supplementary material contains additional facets of `pirouette`, such as the installation of the package, an overview of pirouette's main functions and a guide for users, based on multiple experiments that are shown here as well.

For these experiments, we limited the number of replicates by time, aiming at a duration of 24 hours per setting, when run on the Peregrine computer cluster of the University of Groningen. Due to this, for example, a run of 40 taxa only has few replicates, because one run takes 4 hours. For all experiments, the intermediate results can all be downloaded from their respective websites, which is approximately 5 gigabyte in total.

All the figures shown in this section are shown without any aesthetical modifications, with the exception that the arrangement of the sub-figures in subsection 3.5.10, where we aligned parts of the figure by hand.

Here is an overview of the various sections:

- subsection 3.5.1: guidelines for users
- subsection 3.5.2: installation
- subsection 3.5.3: resources, such as website, tutorials, packages used, bug reporting and contributing
- subsection 3.5.4: citation of pirouette
- subsection 3.5.5: the twinning process
- subsection 3.5.6: candidate models for the inference
- subsection 3.5.7: the effects of stochasticity
- subsection 3.5.8: the nLTT statistic
- subsection 3.5.9: main functions
- subsection 3.5.10: code, extra figures and diagnostics regarding the main example.
- subsection 3.5.11: the result of using multiple trees, as generated by the same stochastic process as the main example
- subsection 3.5.12: the effect of the number of taxa
- subsection 3.5.13: the effect of the DNA alignment sequence length
- subsection 3.5.14 shows the effect when performing inference in the simplest use case
- subsection 3.5.15 shows the effect when performing inference with an under-parameterization
- subsection 3.5.17 shows the effect when the twin alignment is allowed to have a different number of substitutions

- subsection 3.5.18 shows the effect of different mutation rates
- subsection 3.6: Acknowledgments
- subsection 3.7: Data accessibility
- subsection 3.8: Author contributions

3.5.1. GUIDELINES FOR USERS

From the experiments shown below, we composed some rough guidelines. These guidelines should be treated as preliminary results, as the total runtime of these experiments is 'only' 19 days.

- The use of 20 replicates results in decent plots.
- The use of more taxa increases the inference error
- The use of longer DNA sequences decreases the inference error.
- When we do not impose the same number of substitutions between true and twin alignment, we observe a difference in the error distributions with respect to the standard case (presented in the main text) where they are forced to have the same number of substitutions.
- Using a mutation rate less than 1.0 / crown age, decreases the inference error. We predict this will increase the error in the parameter estimation.

3.5.2. INSTALLATION

`pirouette` will be made available on CRAN from which it can then be easily installed:

```
install.packages("pirouette")
```

Until it is on CRAN, and for the most up-to-date version, one can download and install the package from `pirouette`'s GitHub repository. We first need the `mcbette` and `nodeSub` packages:

```
remotes::install_github(
  "richelbilderbeek/mcbette"
)
remotes::install_github(
  "thijsjanzen/nodeSub"
)
```

Now we can install `pirouette`:

```
remotes::install_github(
  "richelbilderbeek/pirouette"
)
```

which also installs its dependencies from CRAN.

To start using pirouette, load its functions in the global namespace first:

```
library(pirouette)
```

Because pirouette calls BEAST2, BEAST2 must be installed. This can be done from within R, using:

```
beastier::install_beast2()
```


For the option to select the best candidate model, pirouette needs the "NS" BEAST2 package [Russel *et al.* 2019]. It can be installed from within R, using:

```
mauricer::install_beast2_pkg("NS")
```

3.5.3. RESOURCES

pirouette is free, libre and open source software available at

http://github.com/richelbilderbeek/pirouette,

licensed under the GNU General Public License version 3. pirouette depends on multiple packages, which are: ape [Paradis *et al.* 2004], assertive [Cotton 2016], babette [Bilderbeek & Etienne 2018], DDD [Etienne & Haegeman 2020], devtools [Wickham & Chang 2016], dplyr [Wickham *et al.* 2019], ggplot2 [Wickham 2009], knitr [Xie 2017], lintr [Hester 2016], magrittr [Bache & Wickham 2014], mcbette [Bilderbeek 2019], nLTT [Janzen 2019], phangorn [Schliep 2011], phytools [Revell 2012], plyr [Wickham 2011a], rappdirs [Ratnakumar *et al.* 2016], rmarkdown [Allaire *et al.* 2017], Rmpfr [Maechler 2019], stringr [Wickham 2017], TESS [Höhna 2013, Höhna *et al.* 2016], testit [Xie 2014], testthat [Wickham 2011b] and tidyr [Wickham & Henry 2019].

pirouette's development takes place on GitHub,

https://github.com/richelbilderbeek/pirouette,

which allows submitting bug reports, requesting features, and adding code. To improve quality, pirouette uses a continuous integration service, has a code coverage of above 95% and enforces the most commonly used R style guide [Wickham 2015].

pirouette's is extensively documented on its website, its documentation and its vignettes. The pirouette website is a good starting point to learn how to use pirouette, as it links to tutorials and videos. The pirouette package documentation describes all functions and liberally links to related functions. All exported functions show a minimal example as part of their documentation. The pirouette vignette demonstrates extensively how to use pirouette in a more informally written way.

The code used in this article and more examples that are periodically tested, can be found at

https://github.com/richelbilderbeek/pirouette_examples.

3.5.4. CITATION OF PIROUETTE

To cite pirouette this article from within R, use:

```
> citation("pirouette")
```

3.5.5. THE TWINNING PROCESS

`pirouette` allows to perform a control measurement, by use of a process we call twinning. This control results in an error distribution that is the baseline error of the pipeline. The difference between the 'true' and 'twin' error distributions is caused only by the mismatch between the true tree model and the tree prior used in the actual inference.

The twinning process, T, encompasses two steps: T_1, that generates a 'twin tree' (Fig. 3.1, 1b) and T_2, which generates a 'twin alignment' (Fig. 3.1, 2b). Both twin tree and alignment will be analyzed in the same way as the true tree and alignment.

We define a phylogeny τ as the combination of branching times $\vec{t}$ and topology ψ, and denote as τ_G the phylogeny produced by a (possibly non-standard) generative diversification model, having branching times $\vec{t}_G$ and topology ψ_G.

The first step (T_1) of the twinning process creates a tree τ_T with branching times $\vec{t}_T$ while preserving the original topology ψ_G:

$$\tau_G = (\vec{t}_G, \psi_G) \xrightarrow{T_1} \tau_T = (\vec{t}_T, \psi_G) \tag{3.1}$$

We chose to preserve the original topology to increase the similarity between the twin to the original tree. This works well in the cases of BD or DD models we consider in our example, because all these models make the same assumption about topology (all topologies are equally likely). However, this might not be suitable for new models that assign different probabilities to trees with the same branching times but different topologies. The default option for the twin diversification model p_T is the standard BD model. `pirouette` has a built-in function to use a Yule model as well. Additionally, a user can specify a function to generate a twin tree from any speciation model, such as, for example, a coalescent model.

It is then possible to use the likelihood function L_T for this diversification model to find the parameters θ_T^* (e.g. speciation and extinction rates, in case of a BD model) that maximize this likelihood applied to the true tree, conditioned on its number of tips n_G:

$$\max[L_T(\theta_T|\tau_G, n_G)] \rightarrow \theta_T^*. \tag{3.2}$$

We use θ_T^* to simulate a number $n_T = n_G$ of branching times $\vec{t}_T$ for the twin tree τ_T, under the process p_T, while preserving the topology. We simulate the new branching times using the TESS package [Höhna *et al.* 2016]. For simplicity, when simulating phylogenies we assumed a sampling fraction of 100%. A different choice might have an effect on model performance.

The second step (T_2) of the twinning process simulates the twin alignment with the same clock model, site model and mutation rate used to simulate the alignment on the true. The twin alignment can be simulated in any user-defined way. `pirouette` provides the option simulate it with the same mutation rate as the true alignment. By default, however, not only the same mutation rate is used, but also the total number of substitutions matches the true alignment. The total number of substitutions is defined as the number of different nucleotides between the (known) root sequence compared to the sequences at the tips.

3.5.6. CANDIDATE MODELS

The user has to specify exactly one standard inference model, but may be unsure which one to pick. To account for this, the user can specify a set of candidate inference models. Each of these candidate inference models is run in an initial, relatively short, analysis; the candidate model with the highest evidence (i.e., marginal likelihood) will then be used in another, longer, inference run, resulting in another error distribution. The evidence for an inference model is estimated by nested sampling [Russel *et al.* 2019], using the `NS` BEAST2 package.

If twinning is used, a candidate model that has the highest evidence for the twin alignment is also used to create the twin error distribution.

3.5.7. STOCHASTICITY CAUSED BY SIMULATING PHYLOGENIES

The goal is to evaluate BEAST2's performance on a non-standard tree model, one must also consider the last source of stochasticity: the different phylogenies a tree model generates. A single phylogeny cannot be considered as fully representative of the model. For this reason multiple phylogenies must be considered (at least 100 independent true and twin trees). If the number of considered phylogenies is high enough, the comparison between the main pipeline's aggregated error distribution and its twin counterpart leads to a fair evaluation of the new tree model with respect to the baseline error.

3.5.8. THE NLTT STATISTIC

The nLTT statistic is the absolute difference between the normalized lineages-through-time plots of two trees. The nLTT statistic is chosen, as it can operate on any two trees (regardless of their crown ages and number of taxa) and its results have a clear range from zero to one. This normalized result makes it possible to compare trees from a distribution of trees from any tree model. The nLTT statistic is not suitable, however, to distinguish between a constant-rate BD model and a family of time-dependent models [Louca & Pennell 2020].

3.5.9. MAIN FUNCTIONS

An overview of `pirouette`'s main functions is shown in Table 3.3. All `pirouette`'s functions are documented, have a useful example and sensible defaults.

Name	Description
`pir_run`	Run `pirouette`
`pir_plot`	Show the `pirouette` results as a plot
`create_pir_params`	Create the `pirouette` parameters
`create_alignment_params`	Create the alignment parameters
`create_twinning_params`	Create the twinning parameters
`create_experiment`	Create one experiment
`create_error_measure_params`	Create the error measurement parameters

Table 3.3 | `pirouette`'s main functions and description.

3.5.10. MAIN EXAMPLE

This subsection describes the pipeline of the main example and its diagnostics in more detail.

The pipeline starts at the top-left panel of figure 3.1 (which is identical to figure 3.2), which is the 'true tree'. The 'true tree' is generated by the diversity-dependent (DD) tree model [Etienne & Haegeman, 2020, Etienne *et al.*, 2012], which is a BD model with a speciation rate that is dependent on the number of species, with (an arbitrarily chosen) crown age of 10 time units and an expected number of 6 tips for an extinction rate of 0.1. The carrying-capacity is set to 6. The initial speciation rate λ_0 is chosen such that the expected number of species in a constant-rate BD model would be equal to the number of tips, which amounts to $\lambda_0 = 0.63$. Note that in the main example, a tree was generated with 5 tips, due to stochasticity in the tree generation algorithm.

From this 'true tree', a 'true alignment' is simulated, using the JC nucleotide substitution model and a strict clock model. The resulting alignment is shown at the center-left of figure 3.1.

From the 'true alignment' the generative inference model is run. Of course, it cannot be the actual (DD) model. Instead, the default BEAST2 inference model is used, which assumes a JC nucleotide substitution model, a strict clock model and a Yule tree model. The resulting posterior trees are shown in the center-left panel of figure 3.1.

From this 'generative true' posterior (center-left panel in figure 3.1), the difference between each of its trees is compared to the 'true tree' (top-left panel), using the nLTT statistic, resulting in the error distribution shown in the bottom-left panel of figure 3.1.

Based on the 'true alignment' (center-left panel), the candidate model with the highest marginal likelihood is determined, from a set of 15 models. The set of models consists of all combinations of all 4 nucleotides substitution models (JC, HKY, TN, GTR), all 2 clock models (strict and relaxed log-normal) and 2 birth-death models (Yule and Birth-Death), except the inference model used as the generative model (JC, strict clock, Yule). The inference model that had the highest evidence (as shown in Table 3.8) was the inference model with a JC nucleotide substitution model, an RLN clock model and a BD tree model. The resulting posterior trees are shown in the second panel of the third row of posteriors in figure 3.1.

From this 'best true' posterior, the difference between each of its trees is compared to the 'true tree' (top-left panel), using the nLTT statistic, resulting in the second error

distribution in the bottom row of figure 3.1.

From the 'true tree' (top-left) we generated a BD twin tree (top-right).

From this 'twin tree', a 'twin alignment' was simulated, using the JC nucleotide substitution model and a strict clock model. The resulting alignment is shown in the center-right panel of figure 3.1.

From the 'twin alignment' the generative inference model is run as well. Also here, the default BEAST2 inference model is used, which assumes a JC nucleotide substitution model, a strict clock model and a Yule tree model. The resulting posterior trees are shown in the third panel of the third row of figure 3.1. From this 'generative twin' posterior, the difference between each of its trees is compared to the 'twin tree' (top-right panel), using the nLTT statistic, resulting in the error distribution shown in the third panel of the bottom row of figure 3.1.

Based on the 'twin alignment' (center-right panel), the candidate model with the highest marginal likelihood is determined, from the same set of 15 candidate models. The inference model that had the highest evidence (as shown in Table 3.9) was the inference model with a JC nucleotide substitution model, an RLN clock model and a Yule tree model (Note that this is different from how the twin tree was generated which was with a BD process and the alignment was simulated with a JC substitution model and strict clock model). However, the extinction rate used in simulating the twin tree was practically 0, thus resembling a Yule process. From the 'twin alignment' this best candidate inference model is run. The resulting posterior trees are shown in the fourth panel of the third row of posteriors in figure 3.1.

From this 'best twin' posterior (fourth in third row of figure 3.1), the difference between each of its trees was compared to the 'twin tree' (top-right panel), using the nLTT statistic, resulting in the fourth error distribution in the bottom row of figure 3.1.

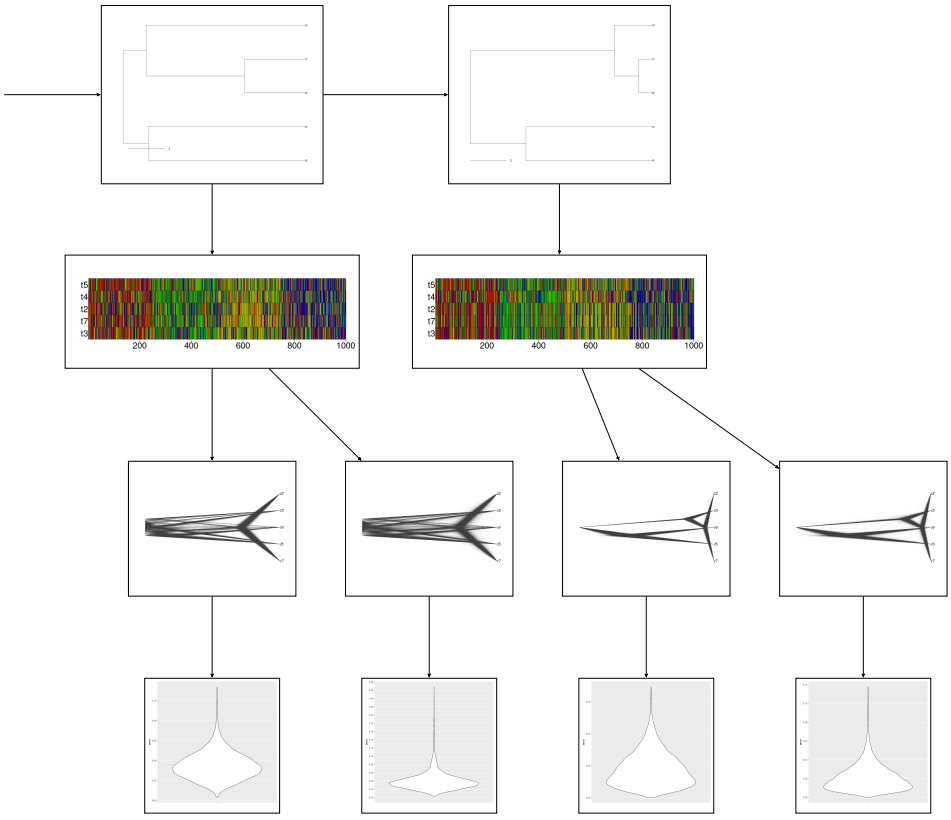


Figure 3.4 | Full pirouette pipeline, including comparison to baseline error. The true tree (top left) is used to simulate an alignment. From this alignment two posterior distributions of trees are created: one using the generative model and another one using the inference model with the highest marginal likelihood. For each distribution of trees, a distribution of errors, measured with the nLTT statistic, between the posterior trees and the main trees is drawn. From the true tree also a twin tree is created (right side of the figure) which follows the same pipeline, leading to two additional error distributions to use as baseline errors.

To assess if the results of the inference are meaningful one important parameter is the Effective Sample Size (ESS). This quantity describes how many independent trees are sampled from the posterior distributions. For reliable results it is good practice to have at least $ESS = 200$ (see

`https://beast.community/ess_tutorial`).

In the following we present the ESS for the posterior distributions of the 4 cases shown in Fig. 3.4.

The ESSes of the 'true' pipeline for the generative model are shown in Table 3.4. From the estimated parameters, one can deduce that the JC nucleotide substitution model was used (no estimated parameter needed), a strict clock model was used (again, no parameter needed to be estimated) and a Yule tree prior is used ('Yule model' and 'birthRate' are estimated). Note that although the actual true tree is created by a DD process, the default and standard Yule tree model is used as the closest standard tree model.

parameter	ESS
posterior	10001
likelihood	10001
prior	9804
treeLikelihood	10001
TreeHeight	10001
YuleModel	9804
birthRate	9931

Table 3.4 | ESSes for generative model

The ESSes of the 'twin' pipeline for the generative model are shown in Table 3.5. Note that the generative inference model is re-used (which assumes a Yule tree model) in the inference, where the twin tree is actually created using a BD process, which is the default.

parameter	ESS
posterior	9969
likelihood	9997
prior	9955
treeLikelihood	9997
TreeHeight	9762
YuleModel	9955
birthRate	9844

Table 3.5 | ESSes for generative model, twin tree

The ESSes of the 'true' pipeline for the best candidate model are shown in Table 3.6. From the names of the estimated parameters, it is clear that the best candidate model has a JC nucleotide substitution model (no parameter needed to be estimated) an RLN clock model (which can be inferred from the parameter 'rate.mean') and a BD tree prior ('BirthDeath', 'BDBirthRate' and 'BDDeathRate').

parameter	ESS
posterior	8320
likelihood	10001
prior	2278
treeLikelihood	10001
TreeHeight	3239
ucldStdev	1027
rate.mean	1853
rate.variance	543
rate.coefficientOfVariation	1215
BirthDeath	7068
BDBirthRate	7615
BDDeathRate	6402

Table 3.6 | ESSes for best candidate model

The ESSes of the 'twin' pipeline for the best candidate model are shown in Table 3.7.

From the names of the estimated parameters, it is clear that the best candidate model for the twin tree is JC nucleotide substitution model (no parameter needed to be estimated), an RLN clock model (which can be inferred from the parameter 'rate.mean') and a Yule model ('YuleModel', 'birthRate'). Note that there is a mismatch between the actual process of how the twin tree and twin alignment are generated, as the twin tree is generated by a BD process, and the alignment is simulated using a JC nucleotide substitution model and a strict clock model. Again we note that the extinction rate used to simulate the twin tree (estimated from the true tree) was practically 0, so the BD process resembled a Yule process.

parameter	ESS
posterior	9623
likelihood	10001
prior	2302
treeLikelihood	10001
TreeHeight	4513
ucldStdev	1414
rate.mean	2923
rate.variance	1560
rate.coefficientOfVariation	1625
YuleModel	7854
birthRate	9636

Table 3.7 | ESSes for best candidate model, twin tree

The marginal likelihood (or evidence) data for the model comparison performed in the 'true' pipeline is shown in Table 3.8. The best (that is, the one with the highest model weight) candidate model assumes a JC nucleotide substitution model, an RLN clock and a BD tree model.

Site model	Clock model	Tree prior	log(evidence)	log(evidence error)	Weight	ESS
GTR	RLN	BD	-6661.105	5.895	0.000	11.422
GTR	RLN	Yule	-6650.211	4.669	0.000	8.311
GTR	Strict	BD	-6656.726	5.304	0.000	11.711
GTR	Strict	Yule	-6656.272	5.567	0.000	7.415
HKY	RLN	BD	-6640.067	4.187	0.001	5.982
HKY	RLN	Yule	-6642.854	4.641	0.000	5.865
HKY	Strict	BD	-6661.308	5.857	0.000	9.510
HKY	Strict	Yule	-6646.973	5.396	0.000	5.643
JC	RLN	BD	-6633.353	2.969	0.945	5.770
JC	RLN	Yule	-6636.447	3.363	0.043	6.941
JC	Strict	BD	-6639.650	4.161	0.002	5.476
TN	RLN	BD	-6640.669	3.778	0.001	6.728
TN	RLN	Yule	-6644.276	4.797	0.000	7.155
TN	Strict	BD	-6638.117	3.277	0.008	5.949
TN	Strict	Yule	-6644.336	3.838	0.000	7.806

Table 3.8 | Evidences for the true phylogeny

The marginal likelihood (or evidence) data for model comparison performed in the 'twin' pipeline is shown in Table 3.9. The best (that is, the one with the highest model weight) candidate model assumes a JC nucleotide substitution model, an RLN clock and a Yule tree model. Note that there is a mismatch between the actual process of how the twin tree and twin alignment are generated, as the twin tree is generated by a BD process, and the alignment is simulated using a JC nucleotide substitution model and a strict clock model. The extinction rate in simulating the BD process (estimated from the true tree) was, however, practically 0, so the BD process resembled a Yule process.

Site model	Clock model	Tree prior	log(evidence)	log(evidence error)	Weight	ESS
GTR	RLN	BD	-5571.348	6.569	0.000	14.874
GTR	RLN	Yule	-5557.738	5.566	0.000	5.352
GTR	Strict	BD	-5547.850	4.386	0.000	5.312
GTR	Strict	Yule	-5553.065	5.664	0.000	6.393
HKY	RLN	BD	-5548.450	3.356	0.000	12.060
HKY	RLN	Yule	-5544.146	3.817	0.000	4.167
HKY	Strict	BD	-5559.866	6.078	0.000	6.766
HKY	Strict	Yule	-5565.050	6.695	0.000	5.779
JC	RLN	BD	-5536.454	2.784	0.040	8.767
JC	RLN	Yule	-5533.284	2.679	0.959	6.058
JC	Strict	BD	-5541.446	3.685	0.000	4.370
TN	RLN	BD	-5557.132	4.844	0.000	9.823
TN	RLN	Yule	-5557.777	4.589	0.000	7.928
TN	Strict	BD	-5545.206	4.229	0.000	9.899
TN	Strict	Yule	-5546.328	3.962	0.000	9.313

Table 3.9 | Evidences for twin phylogeny

3.5.11. USING A DISTRIBUTION OF TREES

This subsection extends the main example, by using multiple (instead of one) trees. These trees are produced by running a DD tree simulation with the same parameters as the main example.

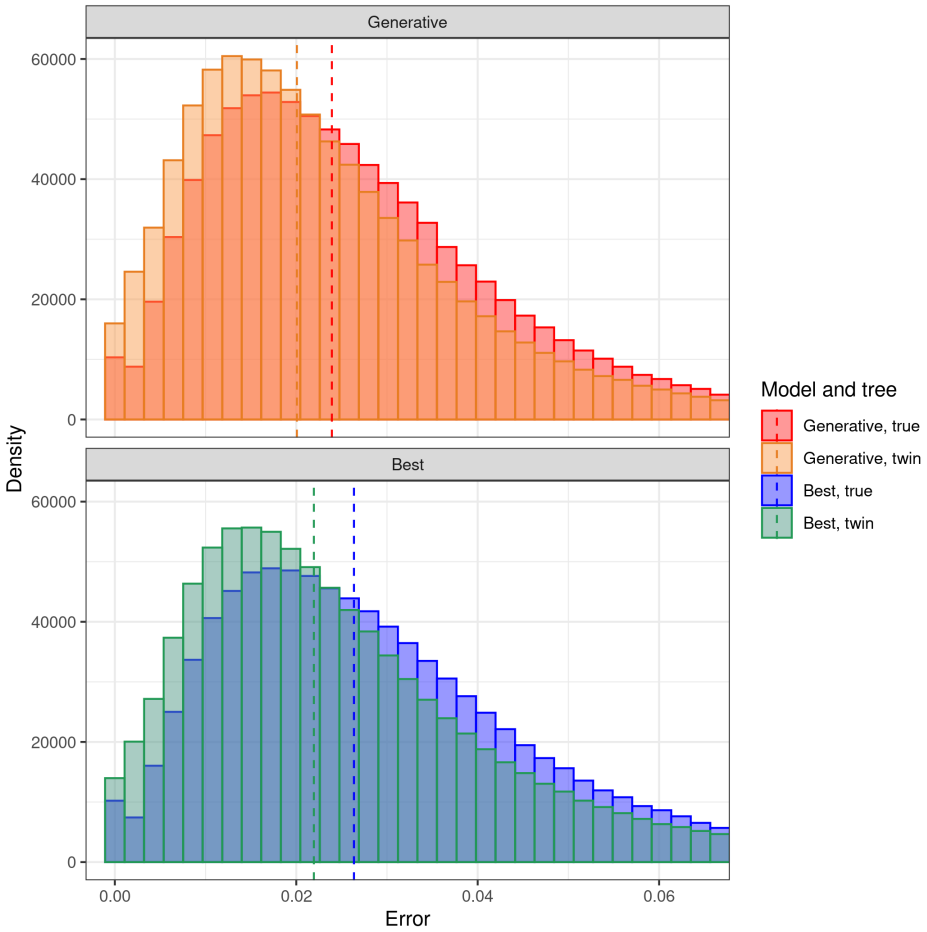


Figure 3.5 | Aggregate error distributions, similar to Fig. 3.3 for the main example, but now for a collection of 100 replicate trees. For each setting (true generative, true best candidate, twin generative and twin best candidate), the resulting errors from each replicate pipeline have been merged into a single distribution. This took 2.7 days (wall clock time) to compute.

The resulting error distributions are shown in Fig. 3.5. We present results for cases where (1) the generative model has been used or (2) the model with highest evidence has been selected for the inference. From the plots we can see that in both cases the two distributions (true and twin) are mostly overlapping, but not everywhere. This suggests that the inference models that have been used can to a reasonable extent capture in

an accurate way the features of the diversity-dependent tree prior used to simulate the original trees.

The code to reproduce Fig. 3.5 can be found at
`https://github.com/richelbilderbeek/pirouette_example_28`.

3.5.12. THE EFFECT OF THE NUMBER OF TAXA

The main example uses 5 taxa. Here we show the same results as the main example, except for a varying number of taxa. We did so, by setting the DD model's carrying capacity to the desired number of taxa.

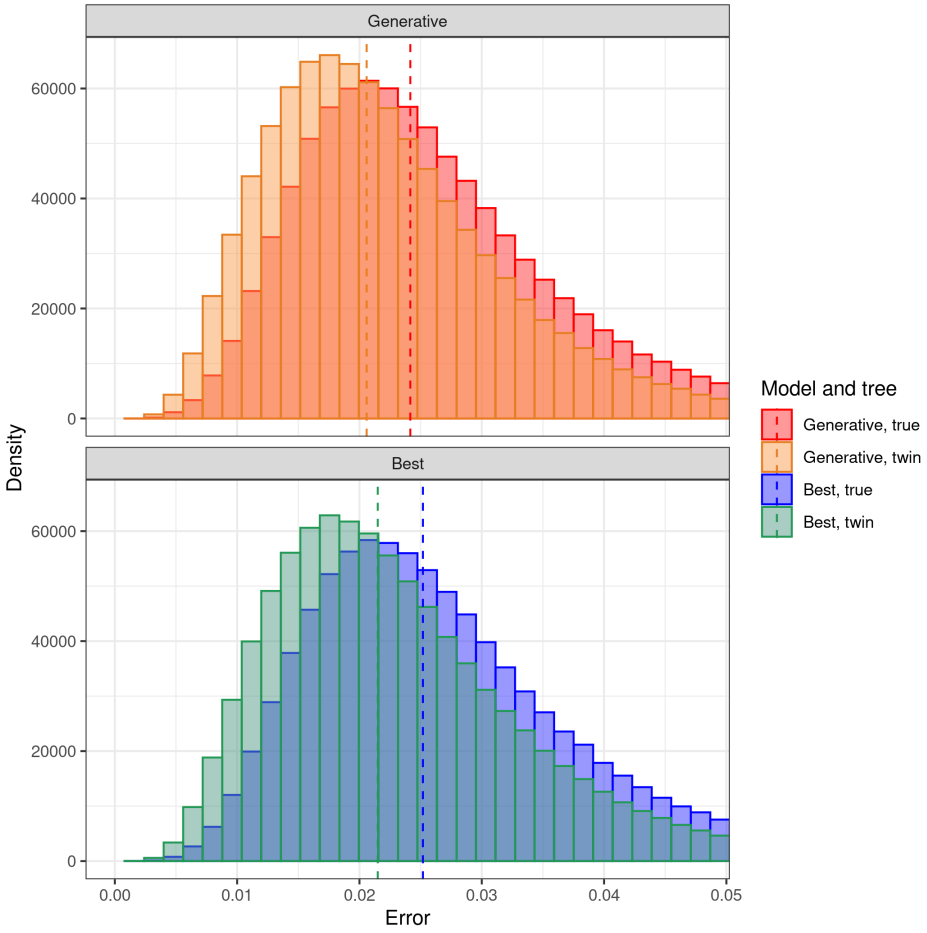


Figure 3.6 | Aggregate error distributions for 100 replicates. Here each true tree has 12 taxa. This took 6.0 days (wall clock time) to compute.

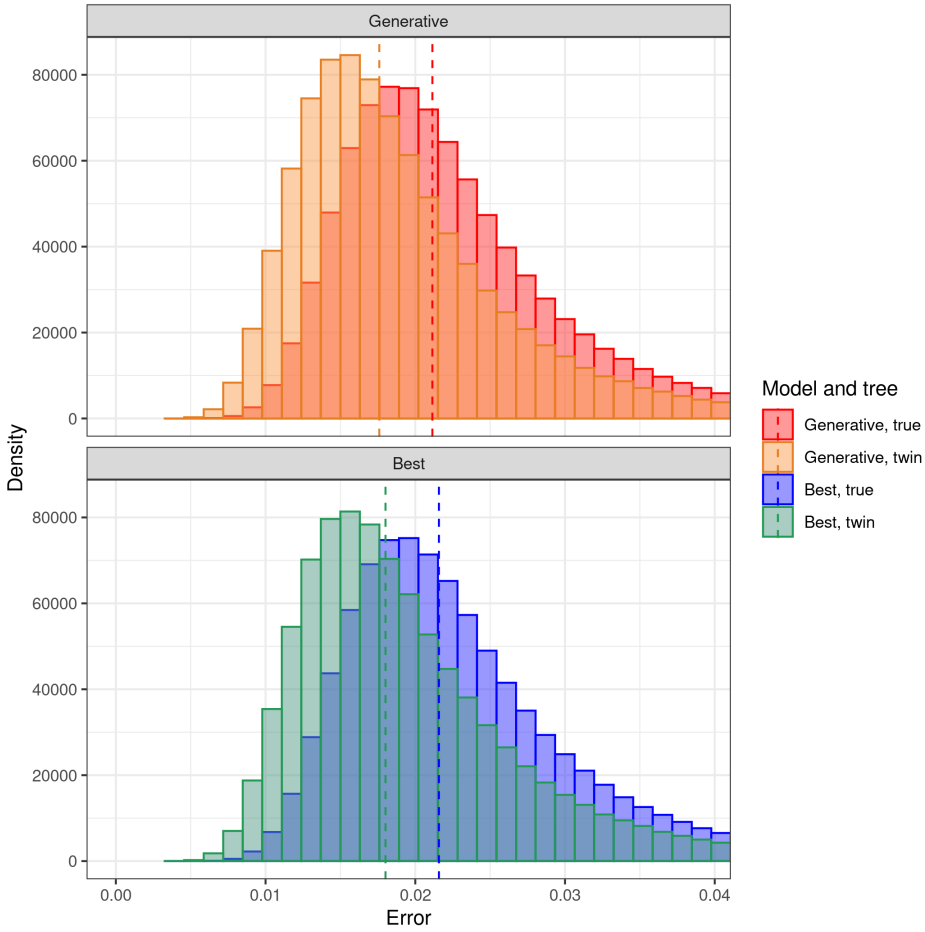


Figure 3.7 | Aggregate error distributions for 100 replicates. Here each true tree has 24 taxa. This took 9.8 days (wall clock time) to compute.

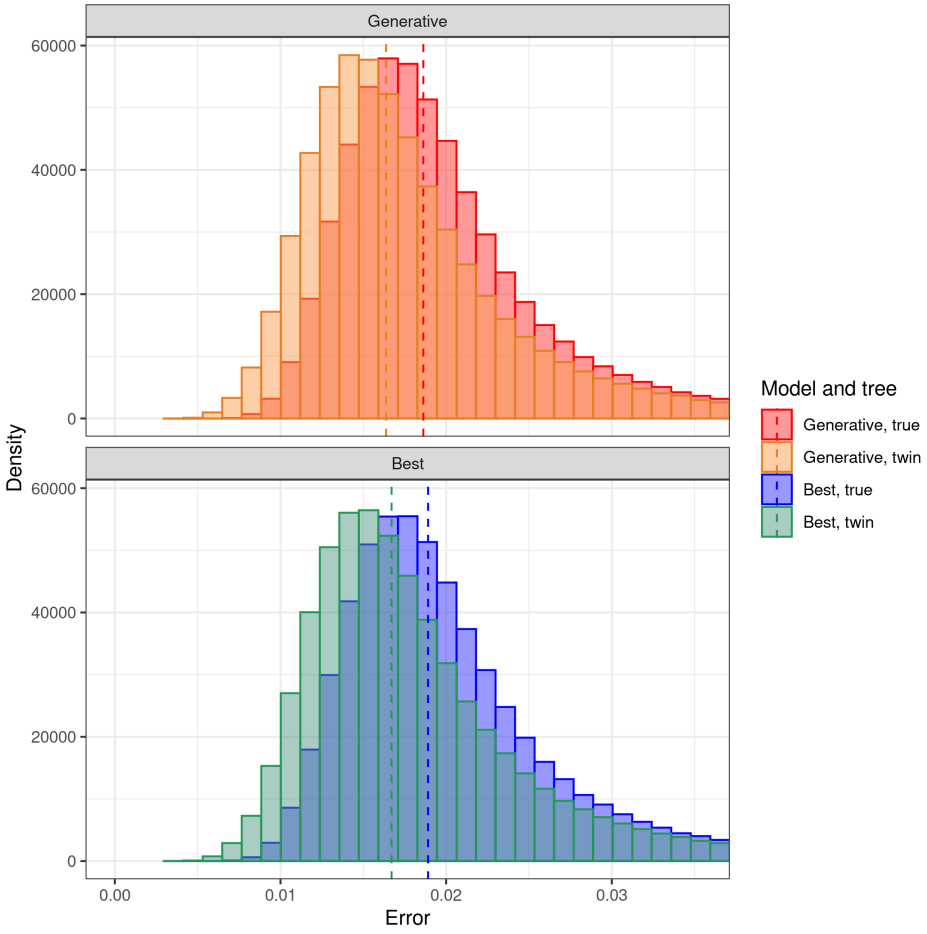


Figure 3.8 | Aggregate error distributions for 65 replicates. Here each true tree has 32 taxa. This took 8.0 days (wall clock time) to compute.

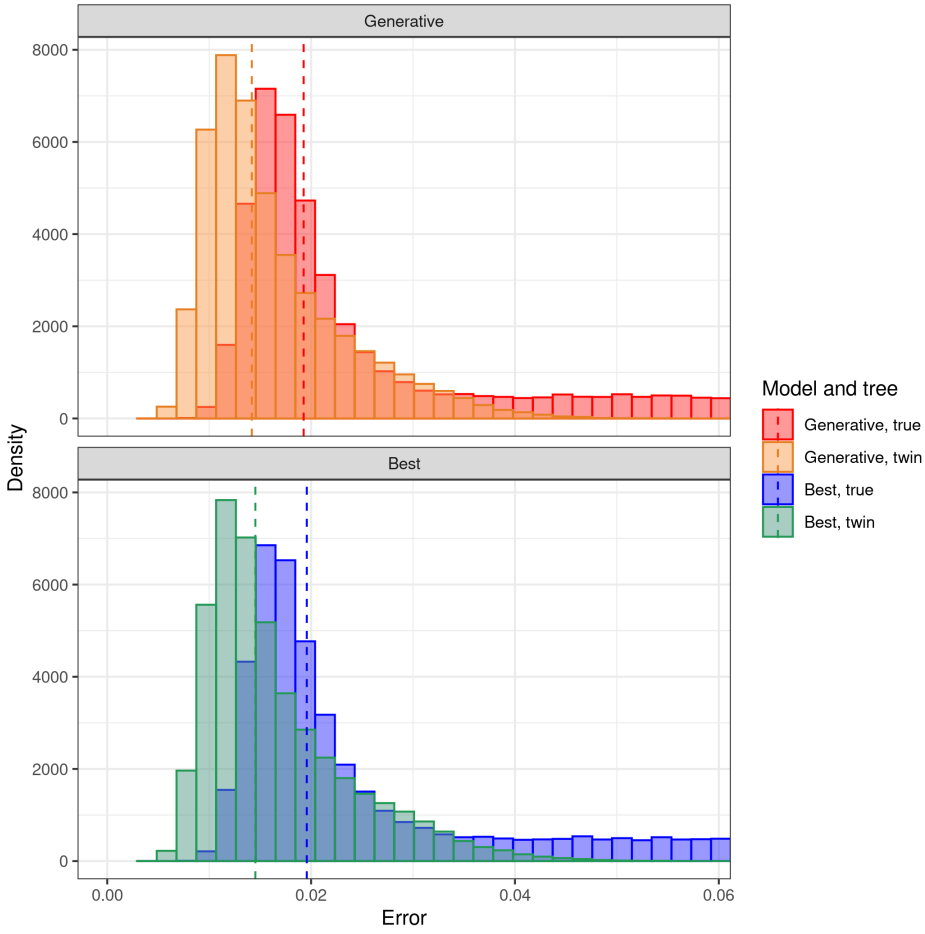


Figure 3.9 | Aggregate error distributions for 5 replicates. Here each true tree has 40 taxa. This took 0.83 days (wall clock time) to compute.

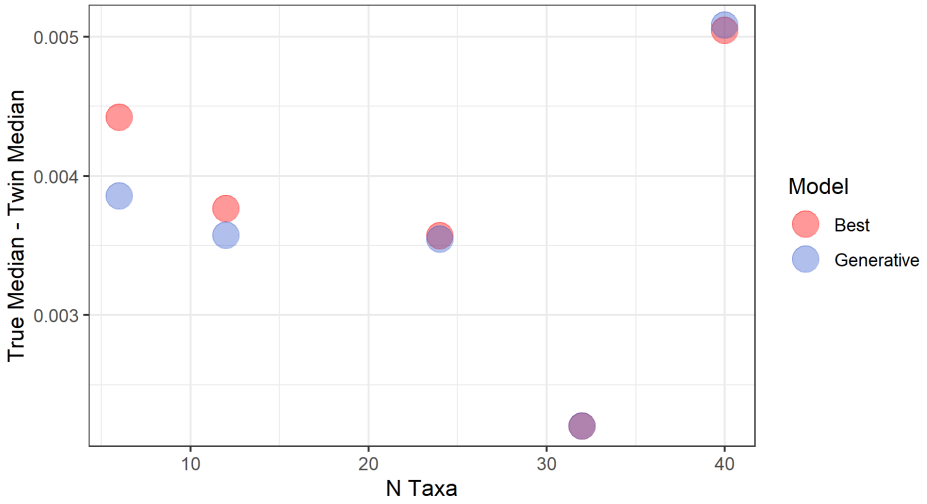


Figure 3.10 | Difference between median true error and median twin error for different number of taxa.

We show in figures 3.5, 3.6, 3.7, 3.8 and 3.9 what are the errors obtained when starting from phylogenies with, respectively, 5, 12, 24, 32 and 40 taxa. Again we can see that in each case errors tend to be greater in the true distribution than in the twin distribution, similar to the result of subsection 3.5.11. Collecting all the data together we can see that errors tend to decrease as the number of taxa in the considered phylogenies increase (see Fig. 3.10). The data point for 40 taxa not following the trend could be due to the limited amount of simulated trees taken in consideration due to time constraints.

The code to reproduce these figures can be found at `https://github.com/richelbilderbeek/pirouette_example_28` (5 taxa, main example), `https://github.com/richelbilderbeek/pirouette_example_32` (12 taxa), `https://github.com/richelbilderbeek/pirouette_example_33` (24 taxa), `https://github.com/richelbilderbeek/pirouette_example_41` (32 taxa), `https://github.com/richelbilderbeek/pirouette_example_42` (40 taxa).

3.5.13. THE EFFECT OF DNA SEQUENCE LENGTH

The main example uses a DNA alignment length of 1000 nucleotides. Here, we show the same results as the main example, except for a varying DNA alignment sequence length.

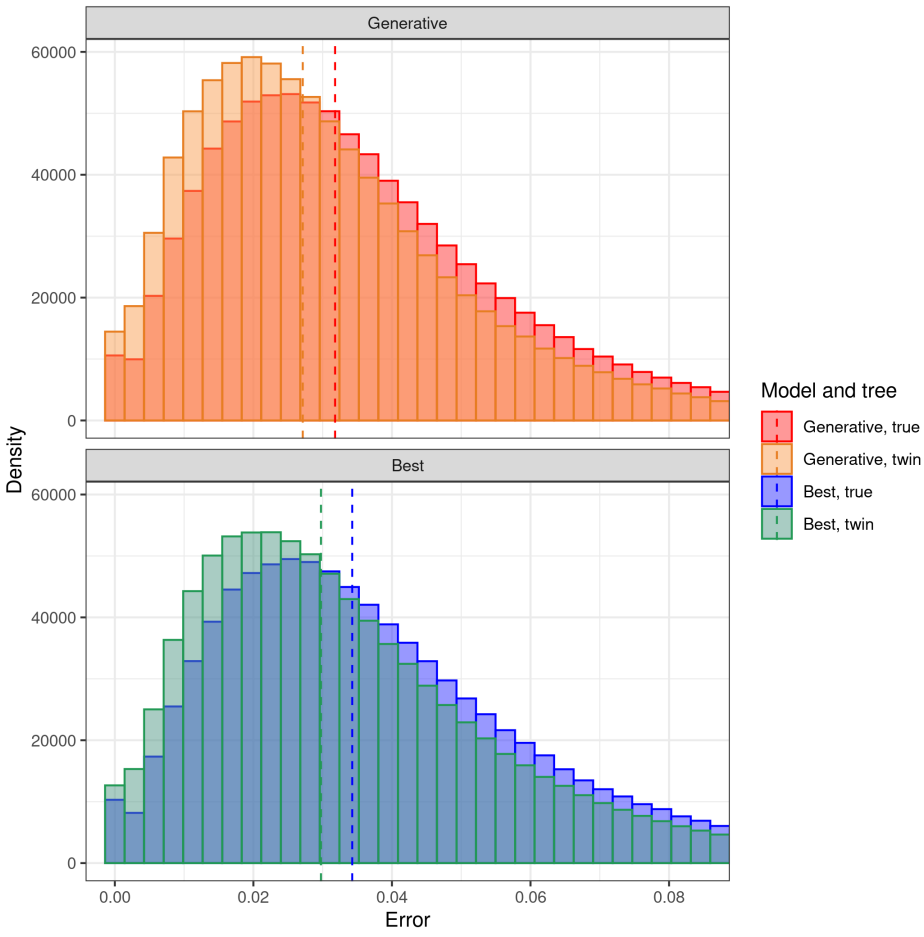


Figure 3.11 | Aggregate error distributions for 100 replicates. Here each each alignment has a sequence length of 500 nucleotides. This took 2.2 days (wall clock time) to compute.

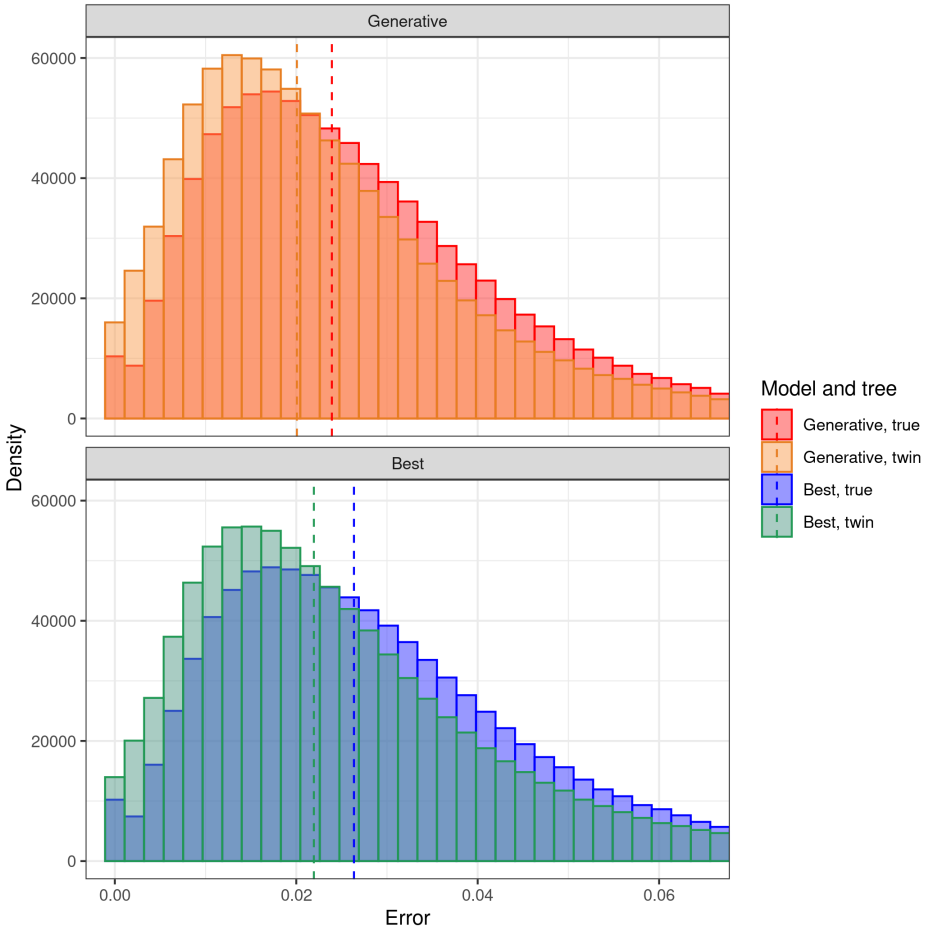


Figure 3.12 | Aggregate error distributions for 100 replicates. Here each each alignment has a sequence length of 1000 nucleotides. This is a replicate of Fig. 3.5. We put it here to facilitate the comparison with the cases with different number of nucleotides.

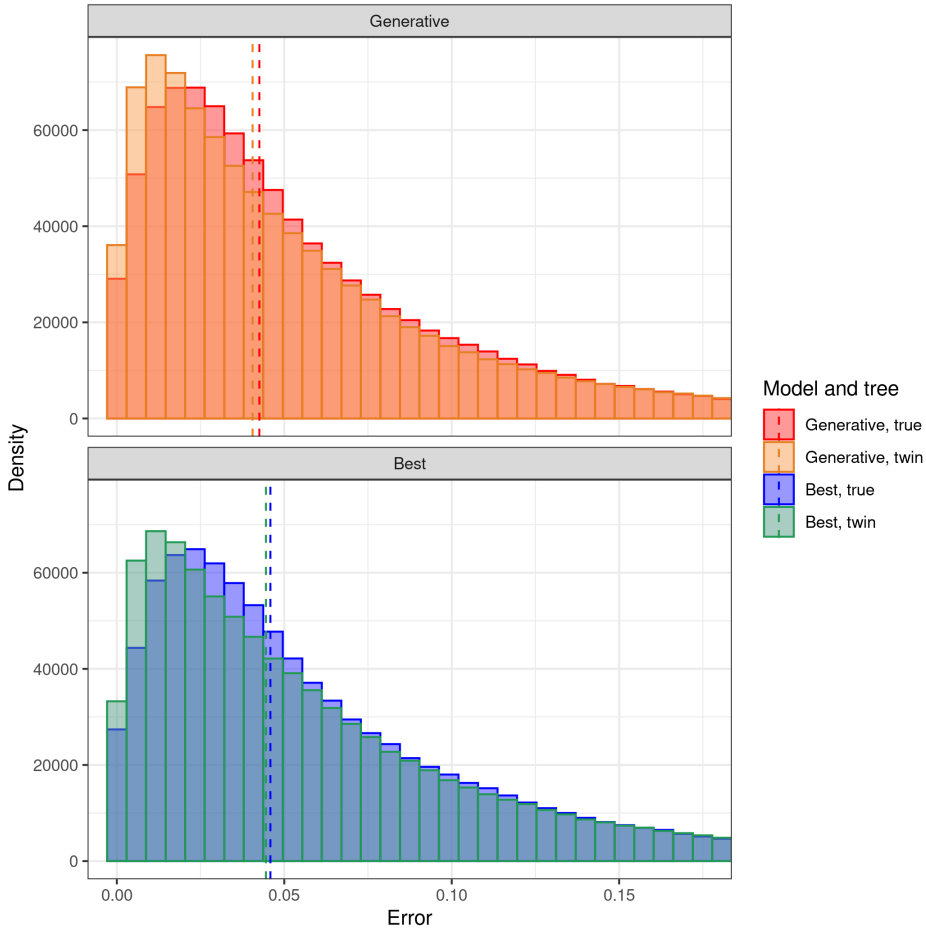


Figure 3.13 | Aggregate error distributions for 100 replicates. Here each each alignment has a sequence length of 2000 nucleotides. This took 4.4 days (wall clock time) to compute.

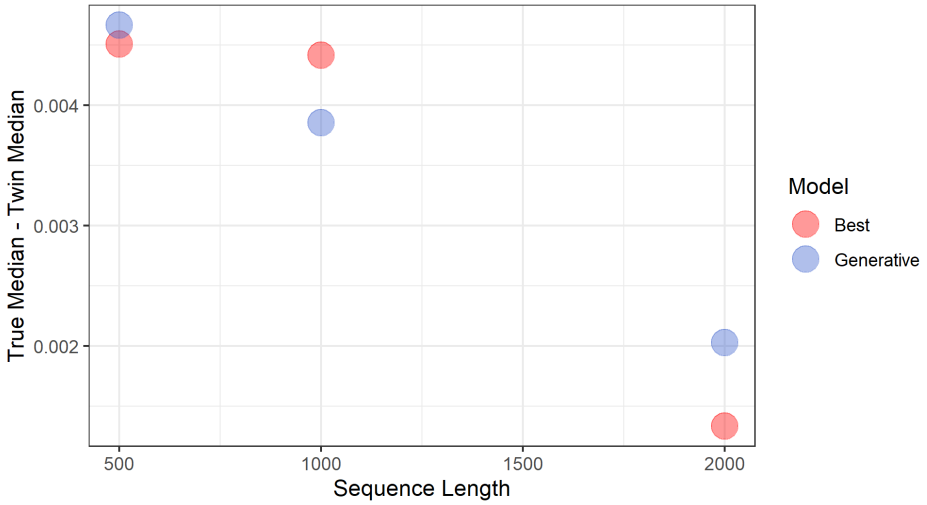


Figure 3.14 | Difference between median true error and median twin error for different sequence lengths.

From figures 3.11, 3.12 and 3.13 we can observe that the discrepancy between the true and twin error distributions tends to become smaller as the number of nucleotides increase (see also Fig. 3.14). This occurred for both the generative and best candidate cases. This follows the expectation that a prior becomes less important when more information becomes available.

The code to reproduce these figures can be found at `https://github.com/richelbilderbeek/pirouette_example_19` (500 nucleotides), `https://github.com/richelbilderbeek/pirouette_example_28` (1000 nucleotides, main example), and `https://github.com/richelbilderbeek/pirouette_example_34` (2000 nucleotides).

3.5.14. THE EFFECT OF ASSUMING A YULE TREE PRIOR ON A YULE TREE

The main example uses a tree generated by a non-standard tree model. Here, we show the same results, with the only difference that the tree used is generated by simplest tree model (the Yule model), which we also assume as the (correct) tree prior.

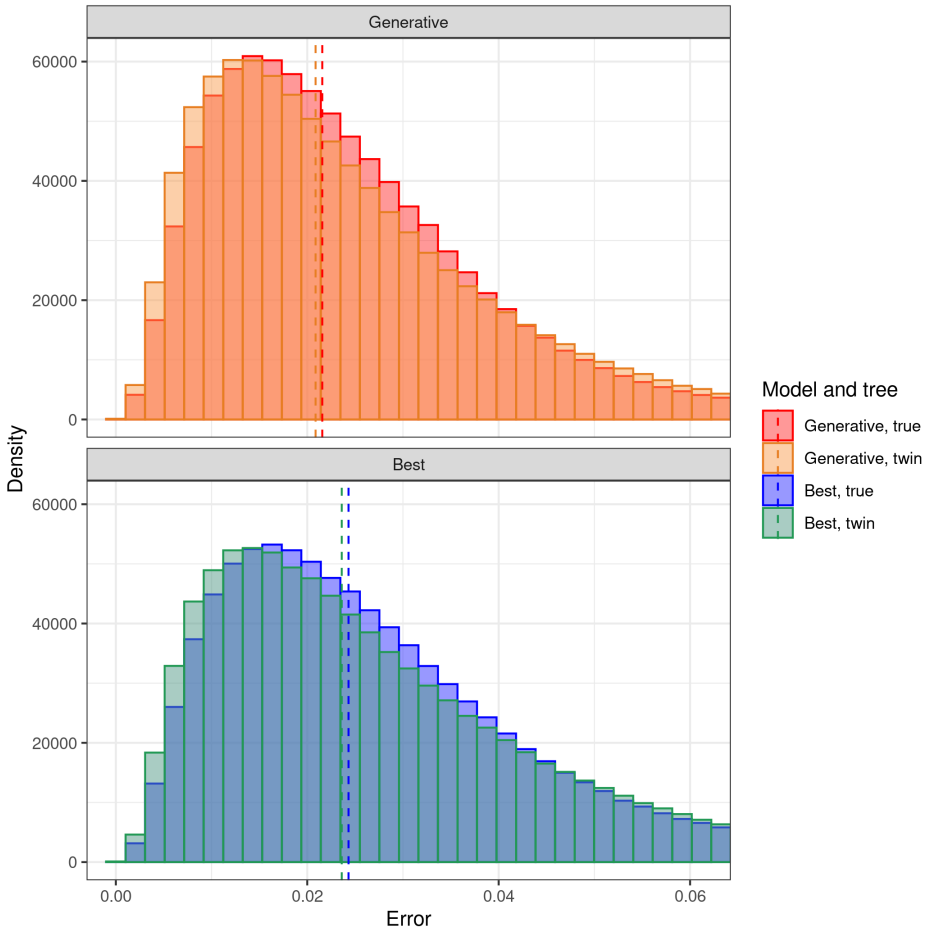


Figure 3.15 | Aggregate error distributions for 100 replicates. Here each true tree is generated by a Yule process. For the inference we used a Yule tree prior. This took 2.9 days (wall clock time) to compute.

This example shows a parameterization at the correct level for the simplest case possible.

As expected the twin and true distributions in Fig. 3.15 are extremely similar for both the generative and the best candidate case.

The code to reproduce this figure can be found at `https://github.com/richelbilderbeek/pirouette_example_22`.

3.5.15. THE EFFECT OF ASSUMING A YULE TREE PRIOR ON A BD TREE

The main example uses a tree generated by a non-standard tree model. Here, we show the same results, with the difference that the tree used is generated by a birth-death (BD) tree model, where we assume it is generated by a Yule (or pure-birth) model. This example thus shows the effect of underparameterization.

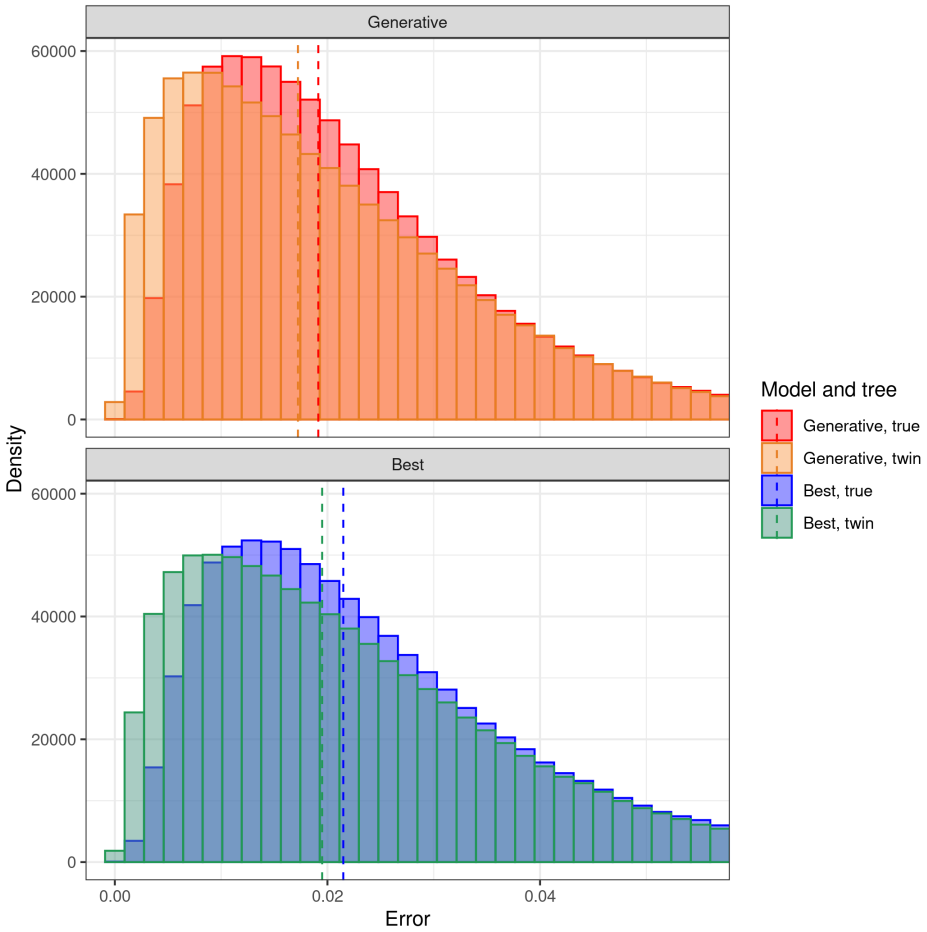


Figure 3.16 | Aggregate error distributions for 100 replicates. Here each true tree is generated by a BD process. For the inference we used instead a Yule tree prior. This took 2.7 days (wall clock time) to compute.

Because the two models are very similar to each other (the BD model can be turned into a Yule model just by setting the extinction parameter to zero [Nee *et al.* 1994]) the median discrepancy is almost negligible. However, with respect to the previous case (subsection 3.5.14), where a Yule tree prior was used, the distributions here exhibit a greater difference.

As we use only extant trees, it is reasonable that the method is slightly weaker in

distinguishing between the Yule and BD models. It is unknown what the discriminatory power would be when comparing trees with extinction events.

The code to reproduce this figure can be found at
`https://github.com/richelbilderbeek/pirouette_example_26`.

3.5.16. THE EFFECT OF DIVERSITY-DEPENDENT TREES DIFFERING IN HOW LIKELY THEY ARE UNDER THE DD PROCESS

Here we show the results of a `pirouette` run on a dataset of multiple DD trees that we selected for having a low, median and high likelihood. In this way, we effectively selected for trees that are rare, uncommon and common respectively.

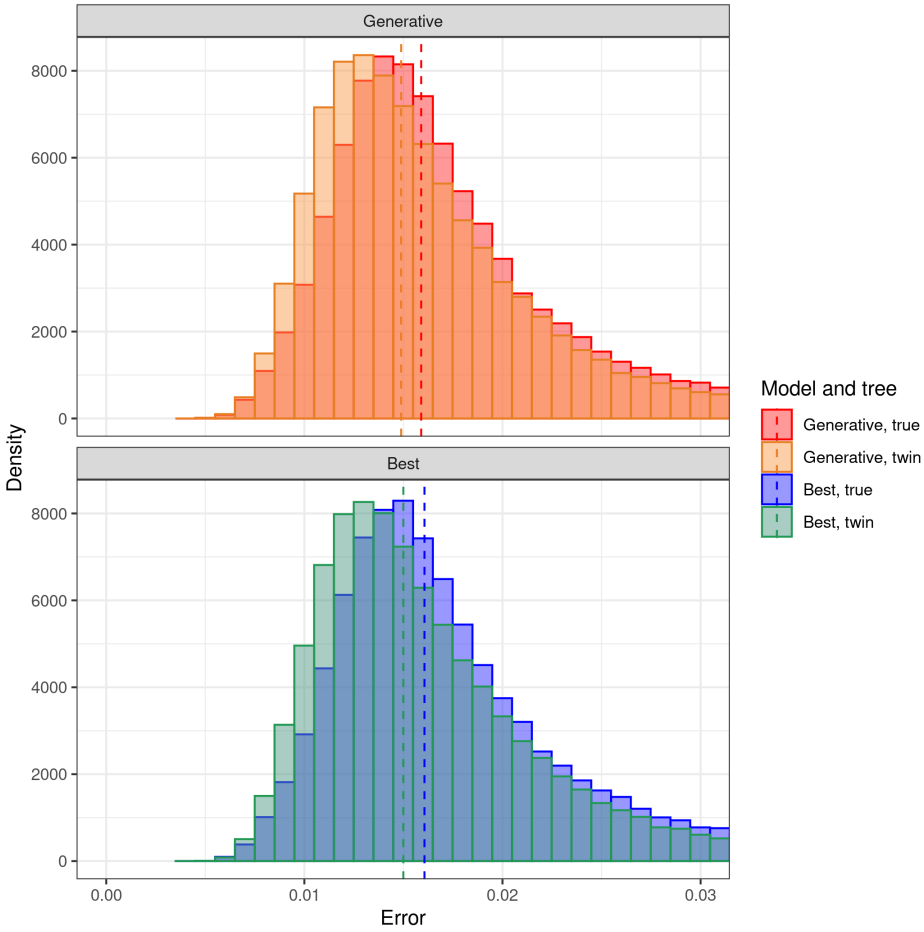


Figure 3.17 | Aggregate error distributions for a distribution of trees, where the true trees are DD with low likelihood.

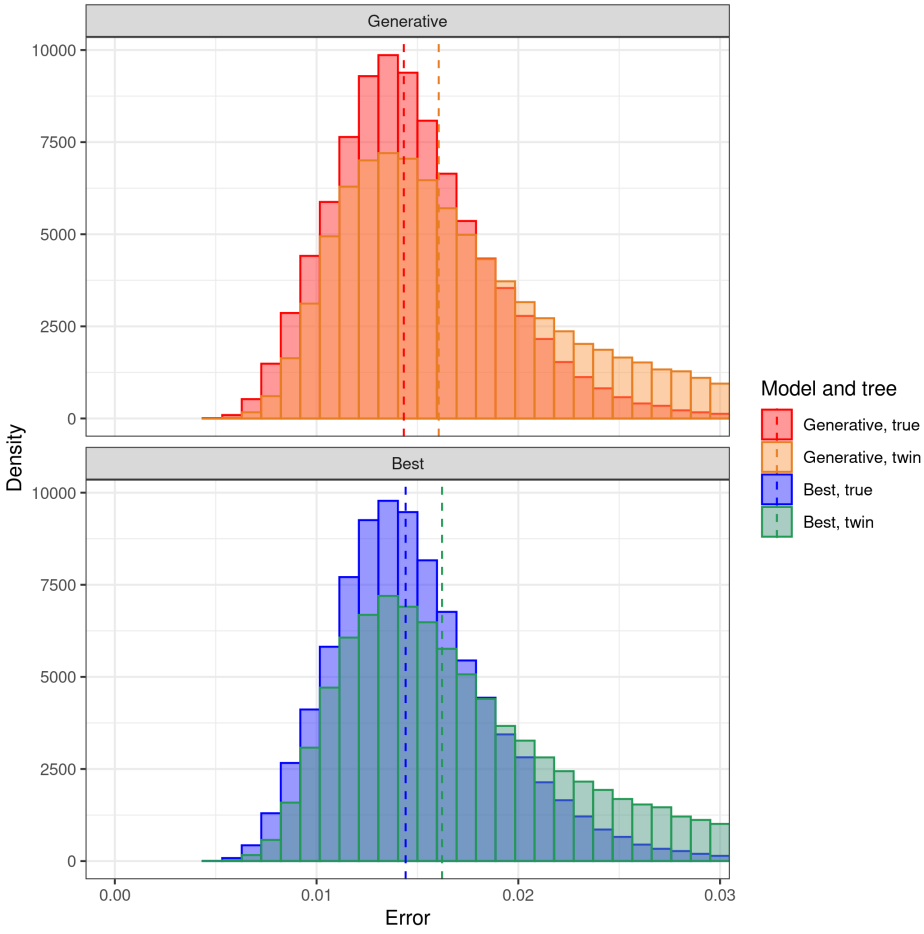


Figure 3.18 | Aggregate error distributions for a distribution of trees, where the true trees are DD with median likelihood.

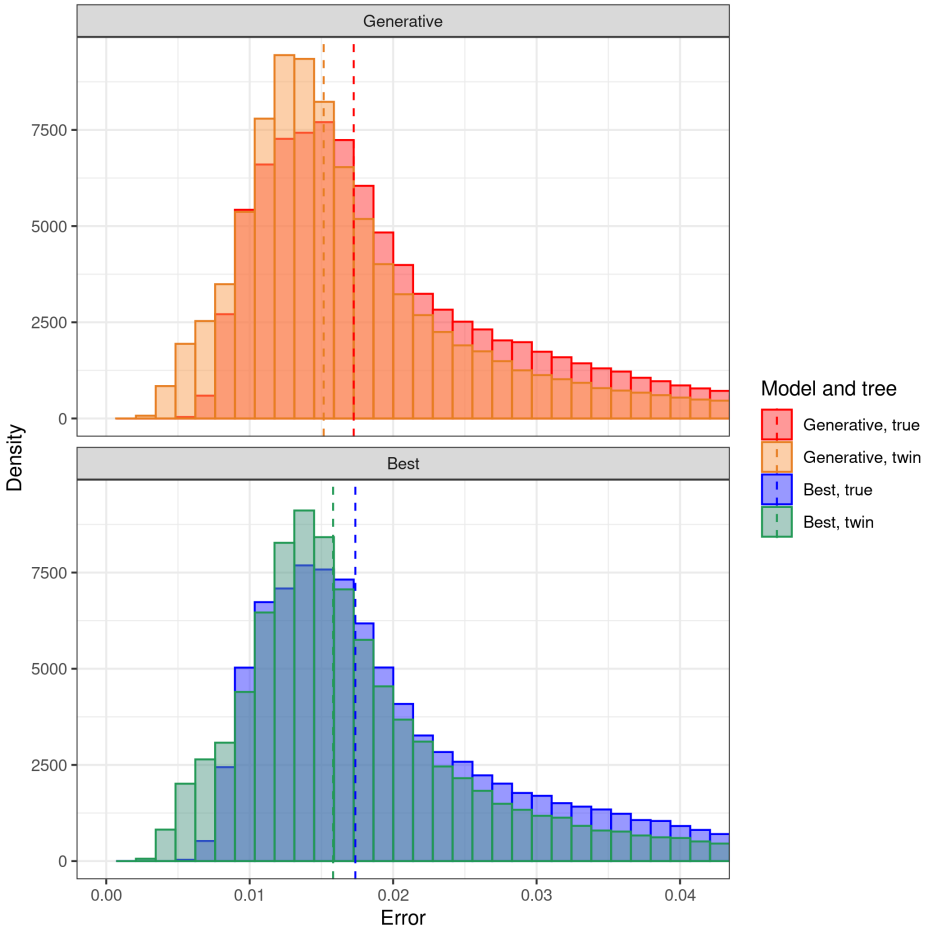


Figure 3.19 | Aggregate error distributions for a distribution of trees, where the true trees are DD with high likelihood.

Here the median errors are similar in the three settings and similar to ones relative to the full dataset of 3.5.11. We can also notice that in the case of median likelihood, the twin median error appears to be lower than the true mean error. This is usually a sign that the number of replicates (in this case 10) is too low to allow us to draw precise conclusions from this test. We did not explore further in this direction using more simulations because computational times turned to be extremely high. The entire run took 120 hours in total.

The code to reproduce these figure can be found at
`https://github.com/richelbilderbeek/pirouette_example_23`
.

3.5.17. THE EFFECT OF EQUAL OR EQUALIZED MUTATION RATE IN THE TWIN ALIGNMENT

The main example uses a twin alignment that has the same number of substitutions (as measured from the ancestral sequence) as the true alignment. Here, we show the same results, with the difference that the twin alignment uses the same mutation rate, yet is not guaranteed to have the same number of substitutions.

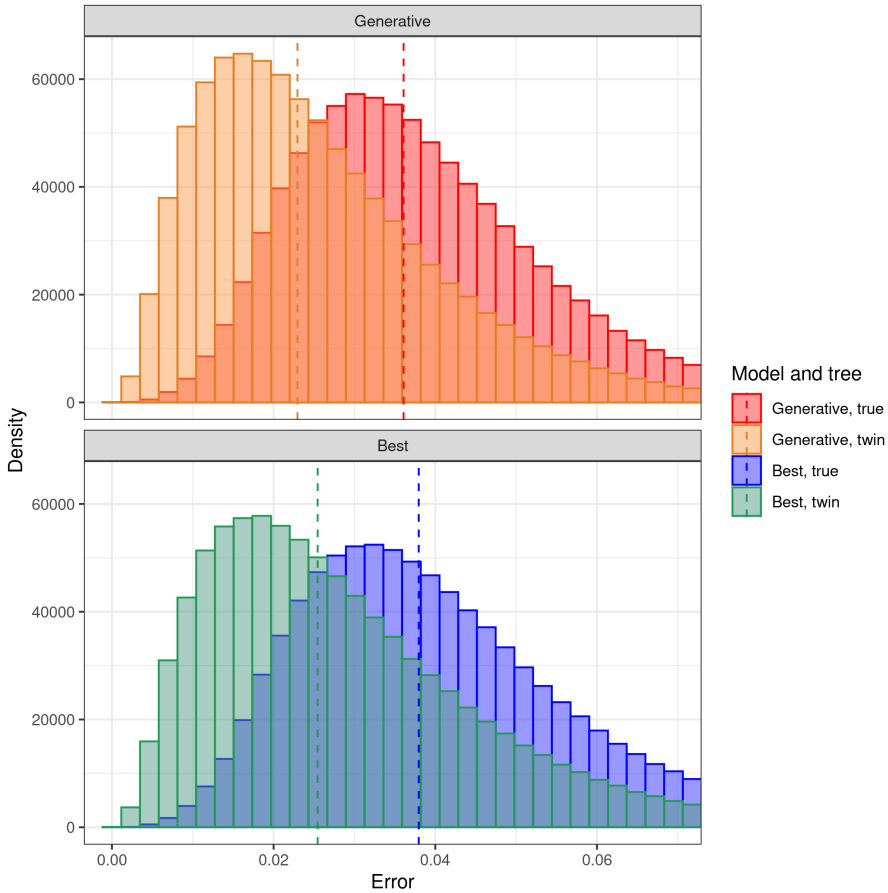


Figure 3.20 | Aggregate error distributions for 100 replicates. similar to Fig. 3.5, but here the number of substitutions is not imposed to be the same between true and twin alignment. Instead, an equal mutation rate is used. This took 3.3 days (wall clock time) to compute.

Comparing figures 3.20 and 3.5 we can see that the discrepancy between true and twin distributions tend to increase. This is probably due to the fact that letting mutation rates induces a difference in the amount of information contained in the alignments and this is reflected in the error distributions.

The code to reproduce this figure can be found at

`https://github.com/richelbilderbeek/pirouette_example_18` and `https://github.com/richelbilderbeek/pirouette_example_28`.

3.5.18. THE EFFECT OF MUTATION RATE

The main example uses a mutation rate such that all nucleotides, on average, mutate once over the history going from the ancestral sequence at the crown to the alignments at the tips. This value equals '1.0 / crown age'. In this way, the alignment is expected to contain the maximum amount of information.

Here, we show the same results for different mutation rates. The results for the different mutation rates are shown in Figs. 3.21 (0.25 / crown age), 3.22 (0.50 / crown age), 3.23 (0.75 / crown age), 3.5 (1.00 / crown age), 3.24 (1.25 / crown age), 3.25 (1.50 / crown age) and 3.26 (2.00 / crown age). Fig. 3.27 summarizes all the other figures showing on the y-axis, for each value of the mutation rate, the difference between the median of the true distribution and the median of the twin distribution. We can observe a general positive trend as the mutation rate increase, even though the value for 1.5 / crown age suggests to take this result with caution. It is possible, however, that a more regular trend could be observed increasing the number of simulations.

The code to reproduce this figure can be found at

`https://github.com/richelbilderbeek/pirouette_example_35` (0.25 / crown age), `https://github.com/richelbilderbeek/pirouette_example_36` (0.50 / crown age), `https://github.com/richelbilderbeek/pirouette_example_37` (0.75 / crown age), `https://github.com/richelbilderbeek/pirouette_example_28` (1.00 / crown age, example reported in 3.5.11, see Fig. 3.5), `https://github.com/richelbilderbeek/pirouette_example_38` (1.25 / crown age), `https://github.com/richelbilderbeek/pirouette_example_39` (1.50 / crown age), `https://github.com/richelbilderbeek/pirouette_example_40` (2.00 / crown age),

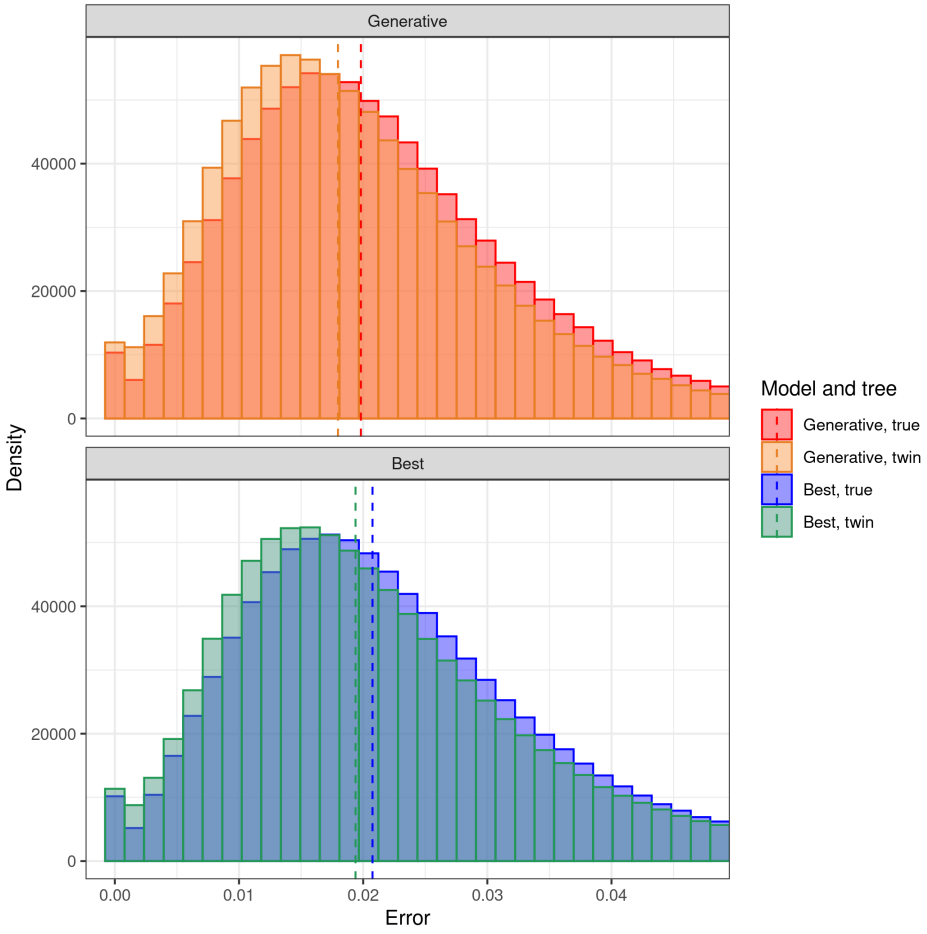


Figure 3.21 | Aggregate error distributions for 100 replicates, for the tree distribution presented in 3.5.11 but with a per-nucleotide mutation rate of 0.25 / crown age. This took 1.8 days (wall clock time) to compute.

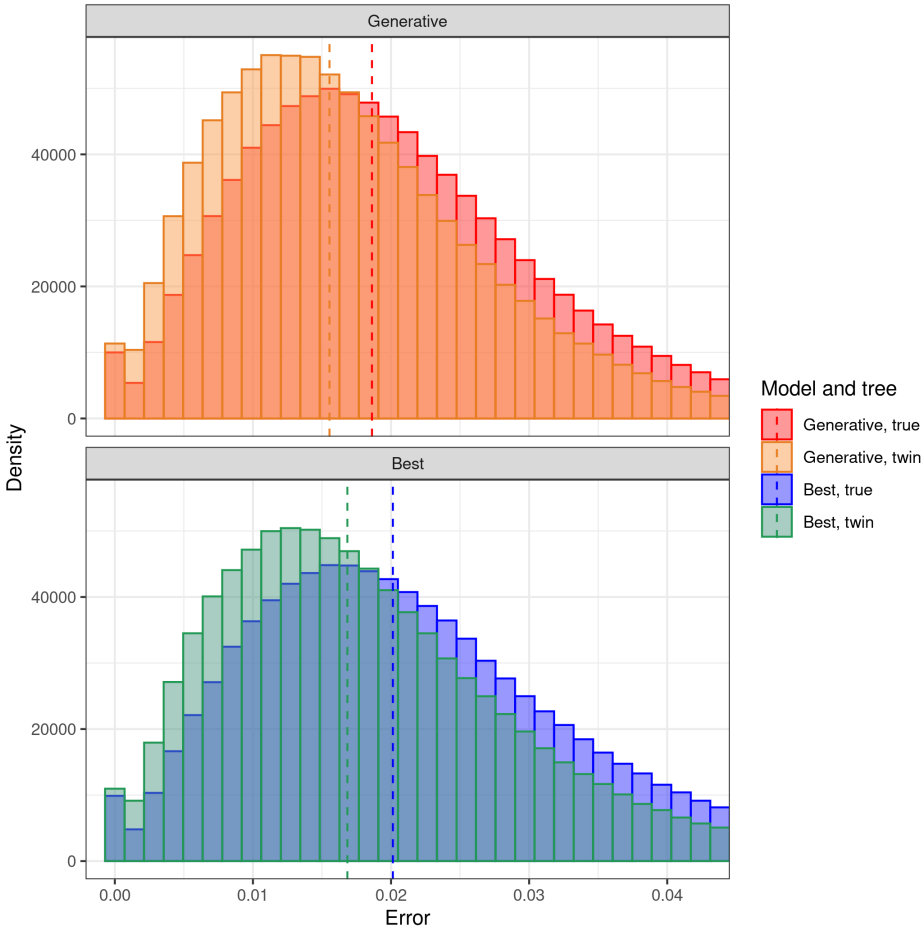


Figure 3.22 | Aggregate error distributions for 100 replicates, for the tree distribution presented in 3.5.11 but with a per-nucleotide mutation rate of 0.50 / crown age. This took 2.2 days (wall clock time) to compute.

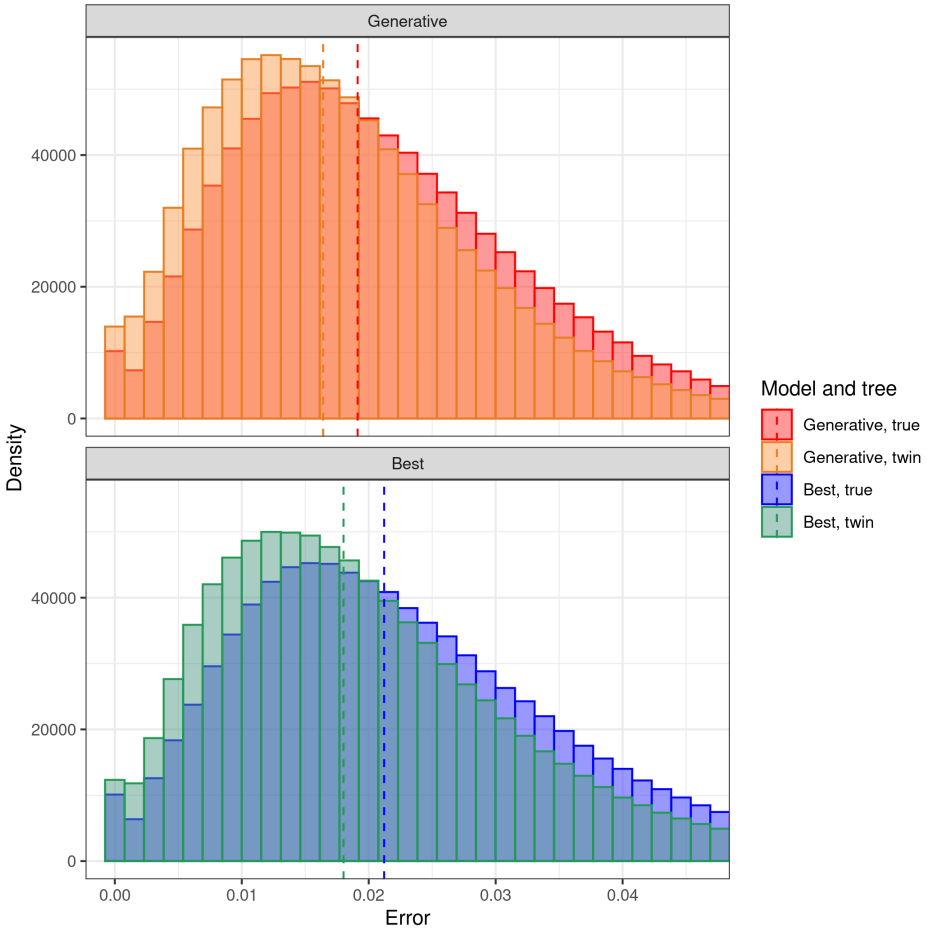


Figure 3.23 | Aggregate error distributions for 100 replicates, for the tree distribution presented in 3.5.11 but with a per-nucleotide mutation rate of 0.75 / crown age. This took 2.5 days (wall clock time) to compute.

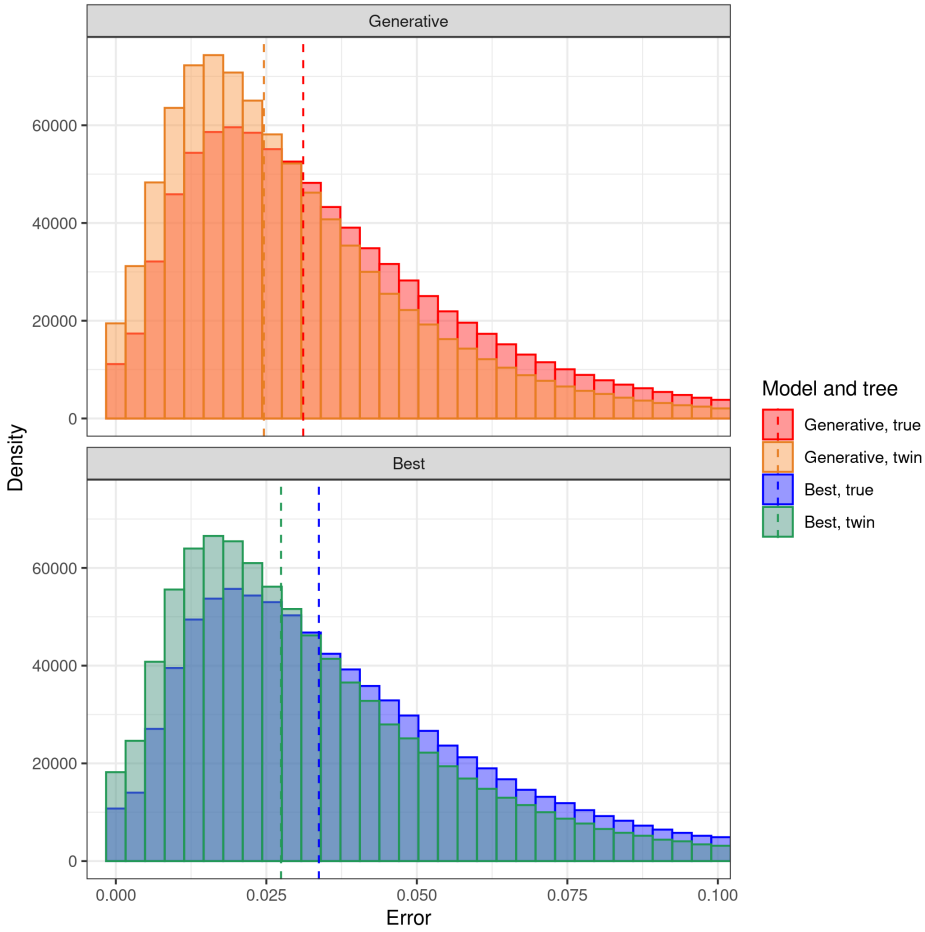


Figure 3.24 | Aggregate error distributions for 100 replicates, for the tree distribution presented in 3.5.11 but with a per-nucleotide mutation rate of 1.25 / crown age. This took 2.9 days (wall clock time) to compute.

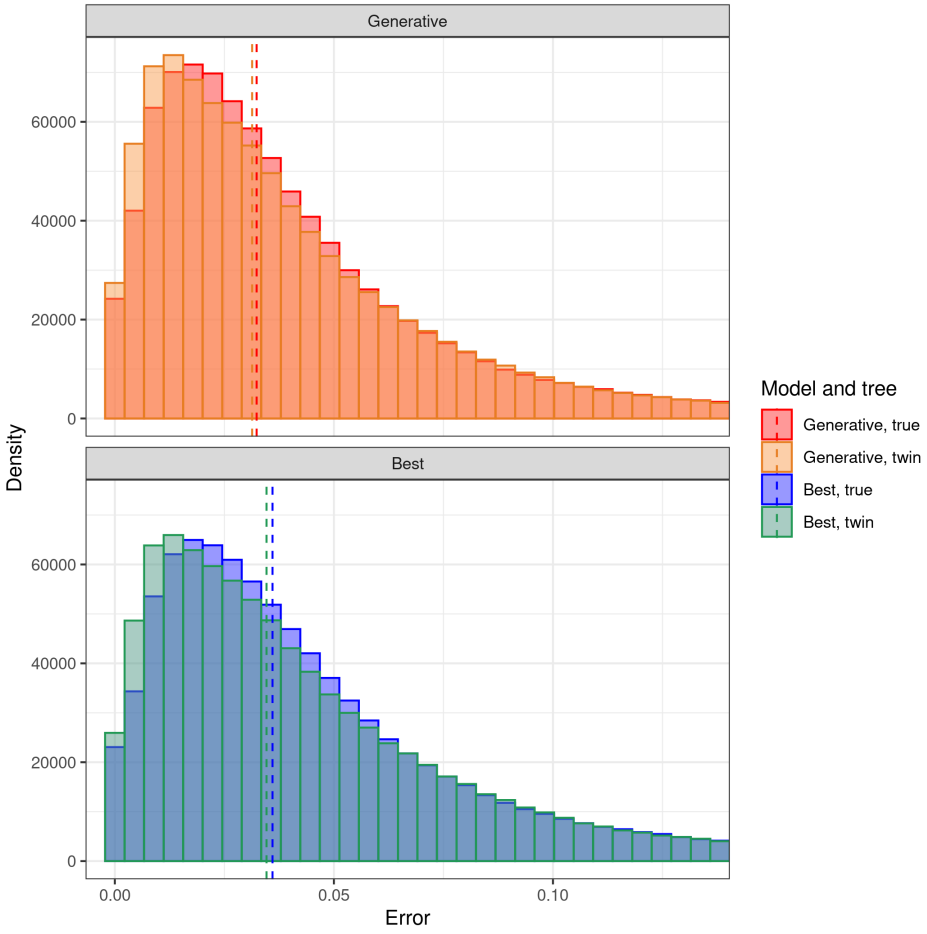


Figure 3.25 | Aggregate error distributions for 100 replicates, for the tree distribution presented in 3.5.11 but with a per-nucleotide mutation rate of 1.50 / crown age. This took 3.0 days (wall clock time) to compute.

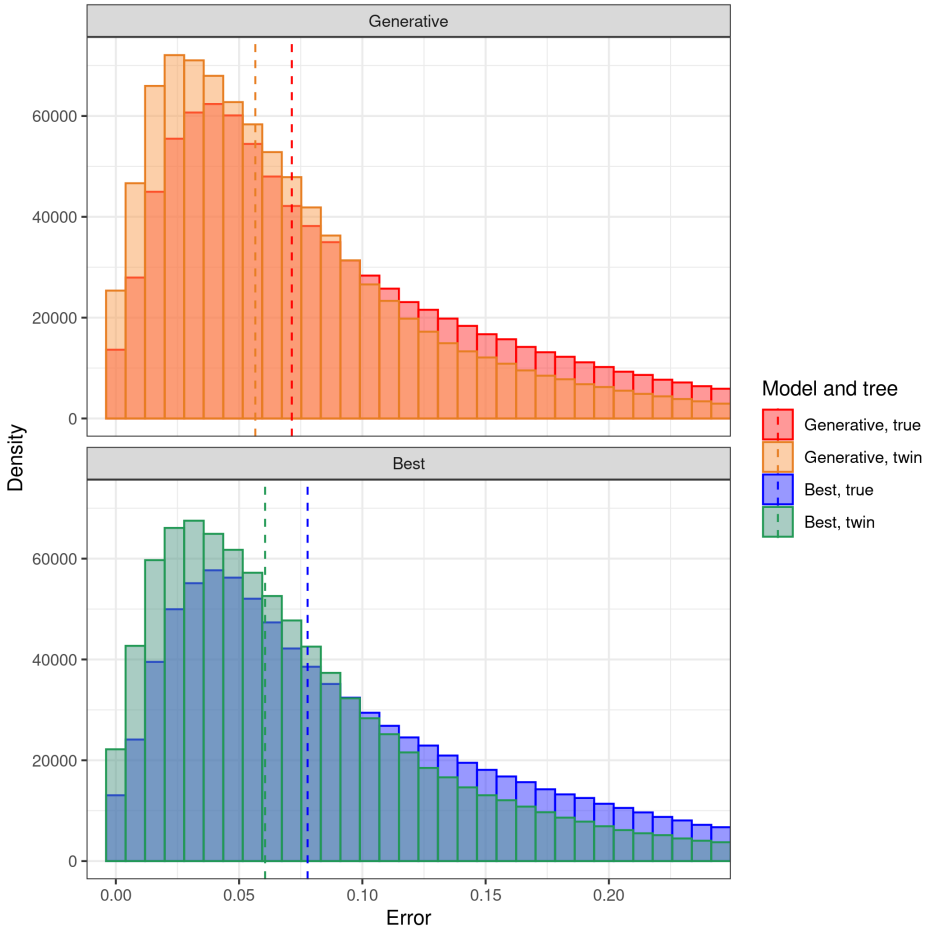


Figure 3.26 | Aggregate error distributions for 100 replicates, for the tree distribution presented in 3.5.11 but with a per-nucleotide mutation rate of 2.0 / crown age. This is done for 100 replicates. This took 3.0 days (wall clock time) to compute.

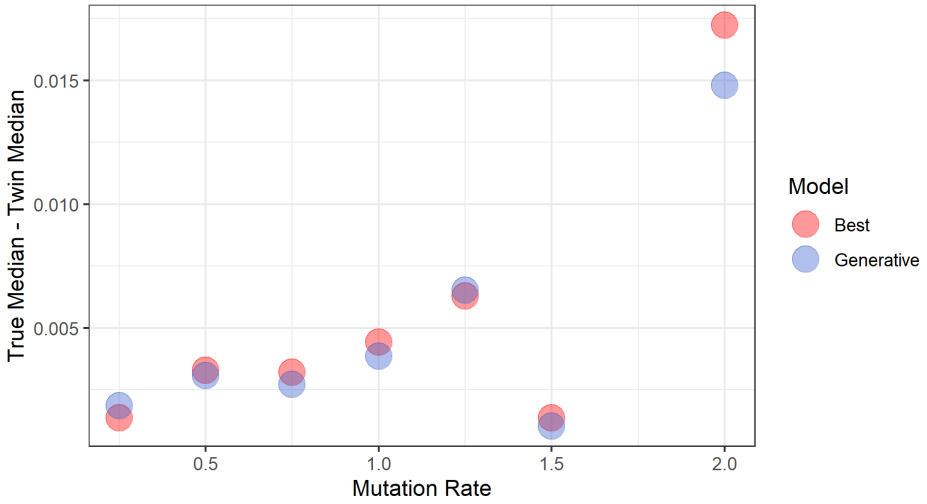


Figure 3.27 | Difference between median true error and median twin error for different values of mutation rate.

3.6. ACKNOWLEDGMENTS

We thank the Center for Information Technology of the University of Groningen for its support and for providing access to the Peregrine high performance computing cluster. We thank the Netherlands Organization for Scientific Research (NWO) for financial support through a VICI grant awarded to RSE.

3.7. DATA ACCESSIBILITY

The pirouette code used for the examples is archived at DOI `https://doi.org/10.5281/zenodo.3969839`. The pirouette examples (including intermediate data) are archived at DOI `https://doi.org/10.5281/zenodo.3970000`.

3.8. AUTHOR CONTRIBUTIONS

RJCB, GL and RSE conceived the idea for the package. RJCB created, tested and revised the package. GL provided major contributions to the package. RJCB wrote the first draft of the manuscript, GL and RSE contributed to revisions.

REFERENCES

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J. & Chang, W. (2017) *rmarkdown: Dynamic Documents for R.* R package version 1.8.

Bache, S.M. & Wickham, H. (2014) *magrittr: A Forward-Pipe Operator for R*. R package version 1.5.

Bilderbeek, R.J. (2019) . `https://github.com/richelbilderbeek/mcbette` [Accessed: 2019-01-21].

Bilderbeek, R.J. & Etienne, R.S. (2018) babette: BEAUti 2, BEAST 2 and Tracer for R. *Methods in Ecology and Evolution*.

Cotton, R. (2016) *assertive: Readable Check Functions to Ensure Code Integrity*. R package version 0.3-5.

Etienne, R.S. & Haegeman, B. (2020) . `https://CRAN.R-project.org/package=DDD`.

Etienne, R.S., Haegeman, B., Stadler, T., Aze, T., Pearson, P.N., Purvis, A. & Phillimore, A.B. (2012) Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 1300–1309.

Hester, J. (2016) *lintr: Static R Code Analysis*. R package version 1.0.0.

Höhna, S. (2013) Fast simulation of reconstructed phylogenies under global time-dependent birth–death processes. *Bioinformatics*, **29**, 1367–1374.

Höhna, S., May, M.R. & Moore, B.R. (2016) TESS: an R package for efficiently simulating phylogenetic trees and performing Bayesian inference of lineage diversification rates. *Bioinformatics*, **32**, 789–791.

Janzen, T. (2019) . `https://github.com/thijsjanzen/nLTT` [Accessed: 2019-04-15].

Louca, S. & Pennell, M.W. (2020) Extant timetrees are consistent with a myriad of diversification histories. *Nature*, **580**, 502–505.

Maechler, M. (2019) *Rmpfr: R MPFR - Multiple Precision Floating-Point Reliable*. R package version 0.7-2.

Nee, S., May, R.M. & Harvey, P.H. (1994) The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London B*, **344**, 305–311.

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

Ratnakumar, S., Mick, T. & Davis, T. (2016) *rappdirs: Application Directories: Determine Where to Save Data, Caches, and Logs*. R package version 0.3.1.

Revell, L.J. (2012) phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, **3**, 217–223.

Russel, P.M., Brewer, B.J., Klaere, S. & Bouckaert, R.R. (2019) Model selection and parameter inference in phylogenetics using nested sampling. *Systematic Biology*, **68**, 219–233.

Schliep, K. (2011) . *Bioinformatics*, **27**, 592–593.

Wickham, H. (2009) *ggplot2: elegant graphics for data analysis*. Springer New York.

Wickham, H. (2011a) The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, **40**, 1–29.

Wickham, H. (2011b) testthat: Get started with testing. *The R Journal*, **3**, 5–10.

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 Packages Easier*. R package version 1.12.0.9000.

Wickham, H., François, R., Henry, L. & Müller, K. (2019) *dplyr: A Grammar of Data Manipulation*. R package version 0.8.1.

Wickham, H. & Henry, L. (2019) *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. R package version 0.8.3.

Xie, Y. (2014) *testit: A Simple Package for Testing R Packages*. R package version 0.4, `http://CRAN.R-project.org/package=testit`.

Xie, Y. (2017) *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.17.

4

RAZZO

Giovanni Laudanno, Richèl J.C. Bilderbeek, Rampal S. Etienne

ABSTRACT

The field of phylogenetics uses heritable material such as DNA to determine the (shared) evolutionary history of a set of species which are summarized in a phylogenetic tree. Bayesian phylogenetic methods allow us to jointly infer probability distributions for the phylogenetic tree and parameters of the various models underlying the methods. One of these models is the diversification model, which mathematically describes the dynamics of speciation addition and removal through time. In Bayesian analyses these are called the (species) tree priors. Tree priors commonly assume that speciation events occur independently. The Bayesian tools heavily rely on this assumption. However, under species pump dynamics, this assumption is violated. By species pump dynamics we mean the (repeated) simultaneous formation of multiple species due to environment-driven isolation which results in temporally aligned (or clustered) divergence times. Current Bayesian phylogenetic tools do not contain such species pump diversification models as tree priors. This may not be a problem if currently implemented tree priors are already capable of inferring a phylogenetic tree to a satisfactory extent. Here we investigate the extent of the error made by one such Bayesian phylogenetic tool (BEAST2) when inferring a phylogenetic tree generated by a known species pump diversification model with a standard tree prior.

To this end we simulate our species pump model, which we call the multiple birth model because it produces multiple simultaneous speciation events, under various parameter settings, and evaluate the corresponding error produced during the inference process. We compare this error with the error made when the generating model is the same as the tree prior used in inference.

We show that the extent of the inference error does not notably increase with the number of multiple birth events. Instead, the phylogenetic inference fails to converge more often under these settings. This reduced convergence is profound and easily detectable, and caused by unknown reasons.

These results show that using standard tree priors for biological systems following a species pump model is warranted, as long as convergence can be attained. For settings that do not converge, the addition of a new species tree prior to the current phylogenetic software may resolve this.

Keywords: computational biology, evolution, phylogenetics, Bayesian analysis, tree prior, pirouette, BEAST2, babette

4.1. INTRODUCTION

Modern computational techniques, such as BEAST (Bouckaert *et al.*, 2014, Drummond & Rambaut, 2007), RevBayes (Höhna *et al.*, 2016) and MrBayes (Huelsenbeck & Ronquist,

2001, Ronquist & Huelsenbeck, 2003), allow to infer phylogenetic trees from genetic data such as DNA, RNA or proteins. They return posterior distributions of phylogenies and estimated parameters by running a Bayesian analysis, given aligned sequence data and a set of models. One of these models is the diversification model, for which a prior distribution must be provided. Within the Bayesian framework this is called a tree prior; it is a mathematical description of the probability distribution of possible branching patterns before looking at the data. Together with the signal from the data, this tree prior will determine the posterior distribution of phylogenies, i.e. after considering the data. Other models include the nucleotide substitution model (i.e. a model of relative transition rates between different nucleotides through time) and the clock model (a model determining the absolute rate of changes for each lineage). For each of these models choices must be made and prior distributions must be specified for their parameters. BEAST2 gives the user the option to set up several possible phylogenetic priors (e.g. substitution/clock/diversification models). However, currently available priors might be not suitable to analyze some specific datasets. For this reason BEAST2 provides users with the possibility to introduce new models and corresponding priors. Particularly, one can specify the tree prior for a new model of diversification.

Current phylogenetic tools such as BEAST2 assume that only a single speciation event can occur at any given time. This assumption is consistent with many different diversification models (e.g Maddison *et al.* 2007, Valente *et al.* 2015, Etienne *et al.* 2012, Etienne *et al.* 2014). However, multiple speciation events can take place simultaneously and repeatedly when populations are intermittently disconnected and connected, for example due to climatic fluctuations. This has been called the species pump hypothesis (Haffer, 1969) and has been invoked particularly in mountainous areas that underwent glaciation (Muellner-Riehl *et al.*, 2019). Our own interest in the species pump hypotheses arose from its potential explanation of the radiation of cichlid fish in the African rift lakes (Malawi, Tanganyika and Victoria), where water level drops created multiple smaller lakes providing the opportunity for allopatric speciation in multiple species. (Verheyen *et al.*, 1996, Sturmbauer *et al.*, 2001, Janzen & Etienne, 2017).

One could study whether the species pump hypothesis is a viable explanation in empirical ystems by comparing divergence times of sister taxa (Oaks *et al.*, 2019). A more inclusive approach would involve using a model allowing multiple simultaneous speciation events as a new species tree prior in phylogenetic reconstruction. However, introducing a new tree prior may be computationally prohibitive (Bilderbeek *et al.*, 2020), and may also not be necessary, as current standard birth death (BD) tree priors might prove to be good enough at inferring the correct tree. Here we use the R package `pirouette` (Bilderbeek & Laudanno, 2019) to check whether this is the case by simulating phylogenies under a species pump model, i.e. the multiple-birth-death model (MBD), with the `mbd` package (Laudanno, 2018), then simulating sequence alignments for each of these trees and finally inferring a phylogenetic tree using BEAST2 from these alignment. By comparing the inferred phylogeny with the true (simulated) one, we measure the inference error made by adopting a standard BD tree prior.

4.2. METHODS

4.2.1. SIMULATION MODEL

The multiple-birth-death (MBD) model inherits the parameters λ and μ from the BD model; they correspond, respectively, to the traditional per-species speciation and extinction rates. Additionally, the MBD model assumes that external events, occuring at rate ν triggers a speciation initiation event in each lineage which leads to a full new species with probability q. Whereas parameter λ can be interpreted as the rate of sympatric speciation, ν is the rate of appearance of geographical barriers able to interrupt the gene flow in the population, resulting in a possible allopatric speciation for each of the species. Even though multiple speciation events can occur simultaneouly, it does not lead to polytomies, because each species can only split once after a trigger event. This model can be easily simulated with a Doob-Gillespie algorithm. A probability distribution for the phylogeny under the MBD model can also be formulated using the integration approach developed for diversity-dependent diversification models Etienne *et al.*, 2012. While this probability distribution could in principle be used as a tree prior in Bayesian phylogenetic inference, it is computationally very demanding, particularly for large trees. With the MBD model we generate simulated datasets for various parameter settings, using the `mbd_sim` function from the `mbd` R package [Laudanno, 2018].

4.2.2. ESTIMATING THE INFERENCE ERROR

From each simulated 'true' MBD tree, we measure the impact of ignoring the more complex and non-standard MBD tree prior in Bayesian phylogenetic inference with the R package `pirouette` [Bilderbeek & Laudanno, 2019].

`pirouette` starts from a 'true' phylogeny (in our case: the simulated MBD tree), and simulates a DNA sequence alignment on it using a known nucleotide substitution model and a clock model. From each sequence alignment, a Bayesian inference is run with a particular choice of tree prior and substitution and clock models. One can choose the same substitution and clock models as used in generating the tree, or pick the ones that fit the data best. For the tree prior we assume the BD model (as the effect of this choice is our focus). We obtain a posterior distribution of jointly-estimated trees and model parameter estimates. By comparing the true tree and the posterior trees, an inference error distribution is generated. For this comparison we used the absolute nLTT statistic by Janzen *et al.* 2015, which results in an error distribution with values ranging from zero (when the inferred tree is identical to the true tree) to a maximum of one (trees are completely different). Another advantage of using the nLTT statistic is that its behavior is best explored in Bilderbeek *et al.* 2020.

We used the twinning option available in `pirouette` that allows to quantify the impact of assuming a wrong tree prior in a Bayesian inference compared to a reference background error that would arise even if the models used in inference were identical to those used in generating the tree (i.e. the twin tree was generated with a BD model). If the error distribution of the true tree matches the error distribution of the twin tree, the effect of using an incorrect tree prior is negligible Bilderbeek *et al.* 2020.

4.2.3. PARAMETER SETTINGS

We ran multiple pilot experiments with increasingly more replicates to arrive at our final parameter settings. We devised a set of rules to make a verdict about the settings.

- quality: 95% of all individual runs should have an effective sample size (ESS) of at least 200, as is recommended by Drummond & Bouckaert 2015.
- feasibility: 95% of all individual runs should finish within 10 days.
- reproducibility: the mean run-time of all finished runs should be less than 24 hours
- relevance 1: the percentage of taxa created by the MB process should be as high as possible
- relevance 2: the percentage of taxa should be as high as possible

We searched through parameter space until these criteria were met. This resulted in the simulation parameters in Table 4.1.

Parameter	Values
λ	(0.2)
μ	(0, 0.15)
ν	(0.0, 0.5, 1.0, 1.5
q	(0.1, 0.15, 0.2)
crown age	8

Table 4.1 | Parameters used to simulate MBD trees. For each parameter setting 40 trees are simulated.

We generated alignments that are 1000 nucleotides in length, with a known root sequence of four 250 mono-nucleotide blocks, generated using the simplest nucleotide substitution model (Jukes Cantor, JC69) and clock model (strict), with a mutation rate of $\frac{0.5}{t_c}$, where t_c is the crown age. With this mutation rate, each nucleotide has a 50% chance to mutate (both silently and non-silently) from the ancestral root sequence to any of the contemporary species' sequences at the tips.

For the Bayesian inference, we assumed a generative model of a site model that follows a JC69 nucleotide substitution model, a strict clock model and a BD tree prior. Additionally, we used a Most Recent Common Ancestor (MRCA) prior equal to the crown age with a normal distribution of width $\sigma = 0.01$. We used a Markov Chain Monte Carlo (MCMC) setup of 10^7 states with a sampling interval of once per 10^4 states. Of the resulting 10^3 states, we discarded a burn-in of 10%.

For each of the MBD parameter settings, we simulated 40 different trees that we put through the same `pirouette` pipeline as described above. We aggregated the error distributions of these 40 replicates.

4.3. RESULTS

Of all runs 98% finished within 10 days. Of these runs, 88% had effective sample sizes above the recommended value of 200, and most well above this (Fig. 4.1). Most runs had less than 100 taxa, of which the median and mean both lie below 30 taxa (Fig. 4.2).

We find no clear difference between the error distributions for the true and twin trees, regardless of whether there is extinction or not in the generating process (Fig. 4.3 and Fig. 4.4. We do observe that for generating values of $\nu \leqslant 0.5$, there is only one mode close to 0, whereas for $\nu \geq 1.0$, the error distributions become bimodal, with a second mode around 0.25. All simulated data can be downloaded from

`https://richelbilderbeek.nl/razzo_project_20200204.zip`.

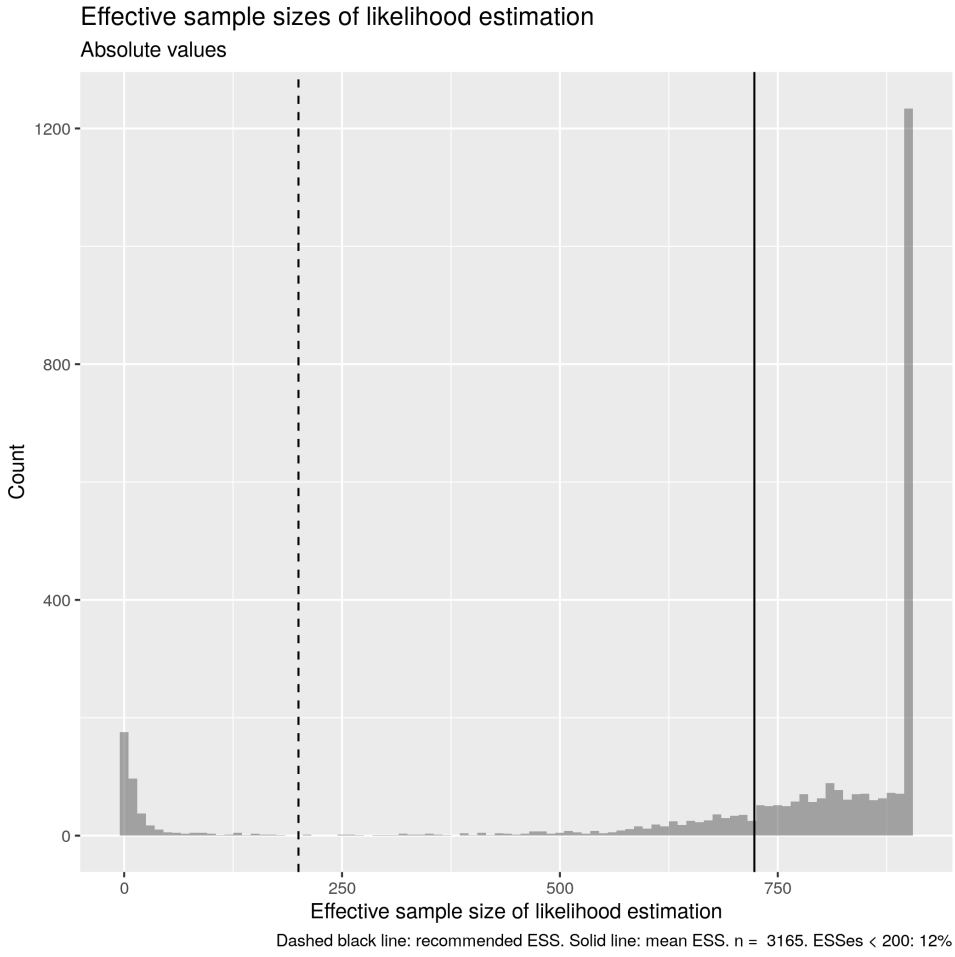


Figure 4.1 | Frequency distribution of effective sample sizes of all experiments that finished within 10 days. Each point represents one parameter setting.

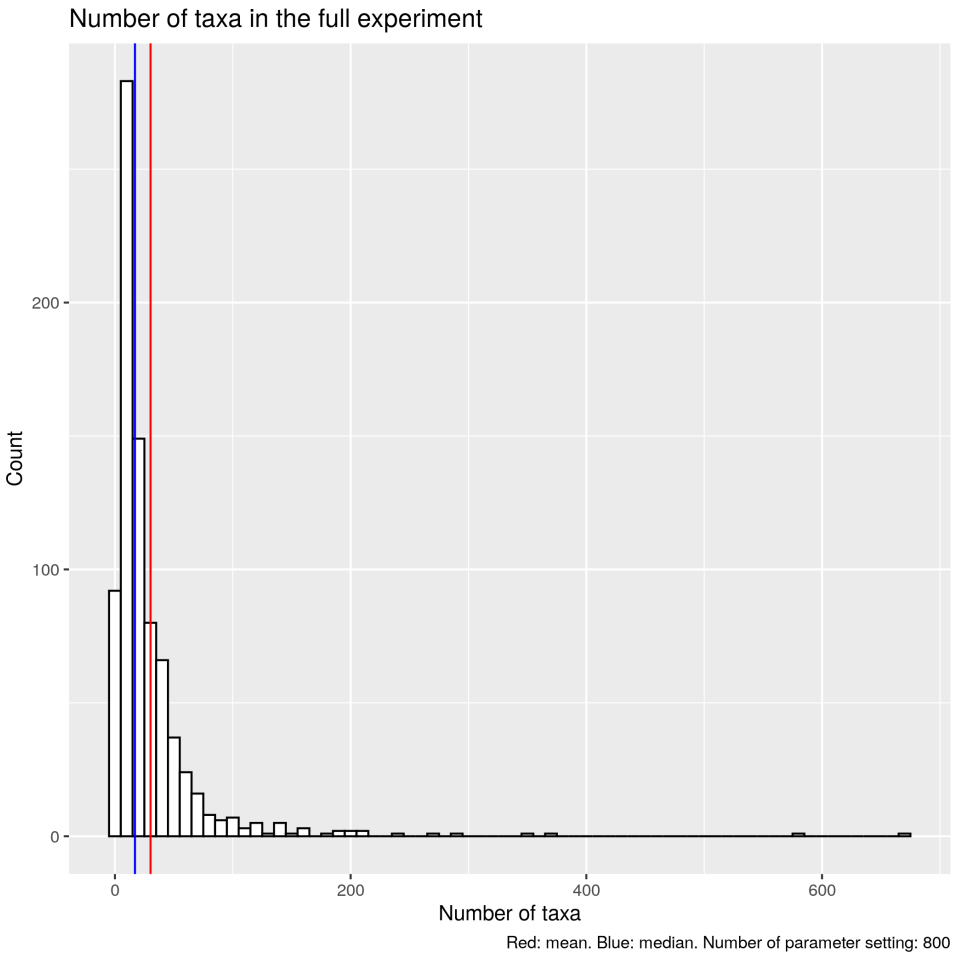


Figure 4.2 | Frequency distribution of the number of taxa across all experiments.

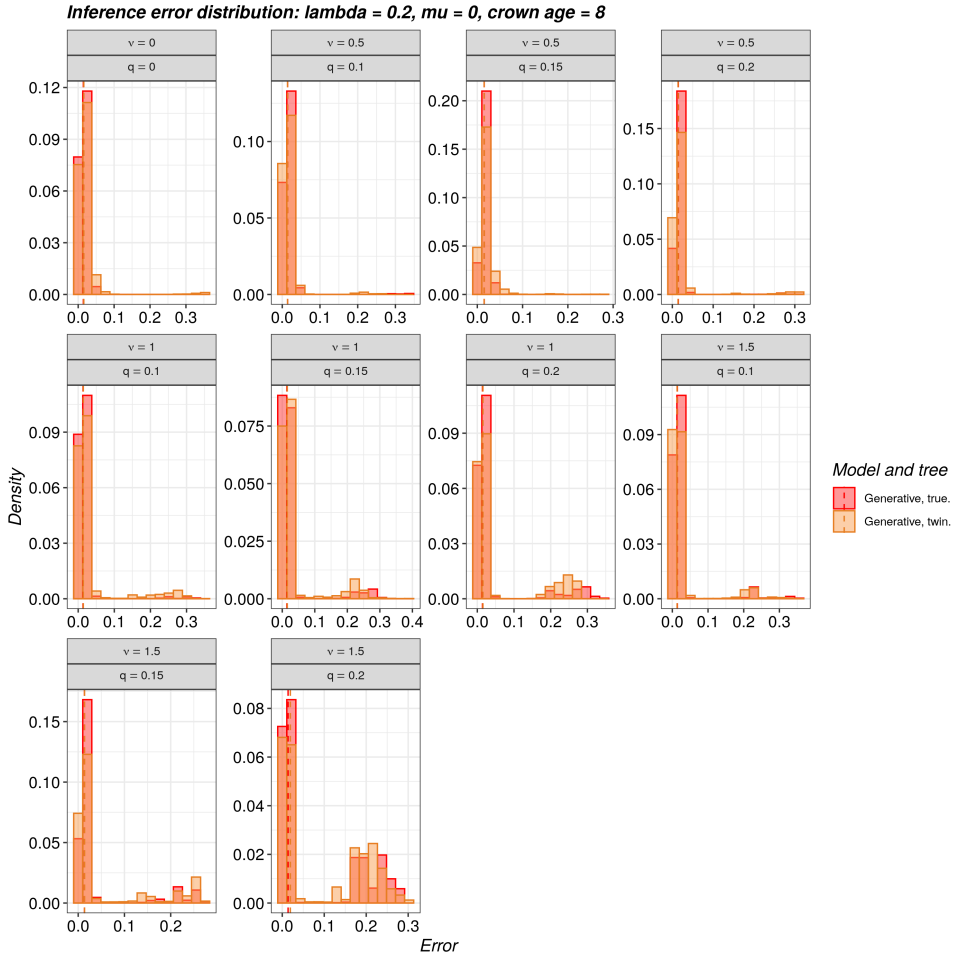


Figure 4.3 | The inference error distribution for a generating extinction rate of $\mu = 0$.

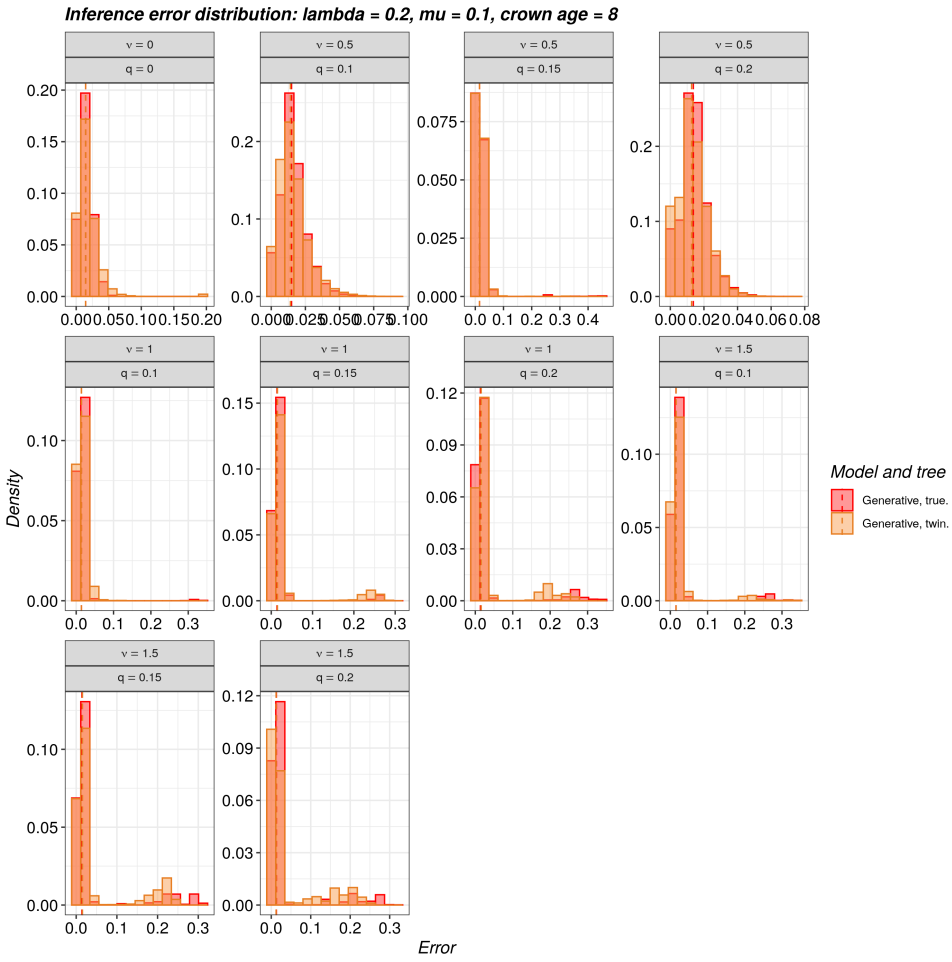


Figure 4.4 | The inference error distribution for a generating extinction rate of $\mu = 0.1$.

4.4. DISCUSSION

We expected to find a larger inference error when increasing the amount of co-occurring speciation events. The reason is simple: the tree prior used assumes that speciation events do not occur simultaneously. The higher the number of co-occurring speciation events, the stronger the deviation from the species tree prior. However, we did not find evidence for this prediction..

The first reason why this may be the case, is because of the noise generated by the runs that did not converge. We think that the bimodality in the error distribution is caused by the converging runs giving a low error, and runs with low ESSes giving higher errors. This can be studied further by Filtering away the runs with a low ESS. However, this will only remove the second mode, but will not explain the similarity of the two error distributions for true and twin trees.

A second reason might be that our choice of MBD parameter settings resulted in too few multiple-birth events. We chose low values of ν to prevent doing inference on huge trees. For larger trees, one might perhaps find a difference in the error distributions. However, we do remark the frequency of multiple-birth events was often larger than 50%.

The third reason why this we did not detect a difference in inference errors between true and twin trees may be due to the choice of measuring our error. We used the nLTT statistic, which may not be sensitive enough to what we consider to be different trees: trees with aligned speciation events and trees where these are not aligned. It is, however, not straightforward to define a metric that is more sensitive to this difference than nLTT. One could perhaps try a shotgun approach by simply applying many commonly used metrics, but we believe that this is unlikely to yield a different outcome. The nLTT statistic at least picks up the sudden increase in number of lineages present in MBD true trees, but not in BD twin trees, but this difference is apparently small relative to other stochastically arising differences between true/twin and the inferred posterior distribution of trees.

The MBD model is a BD model that allows for speciation events taking place simultaneously, but has the same drawbacks as the BD model: the expected number of species increases quickly with time (when net formation of species exceeds extinction) and the model does not take into account that speciation takes time. The MBD model can, at least in the simulations, easily be extended to have a diversity-dependent speciation rate (as in the DDD model) and/or making speciation take time by adding an incipient species state (as in the PBD model). The idea of this experiment, however, was to measure the impact of species pump dynamics on phylogenetic inference that does not assume this, and the comparison of a BD and MBD model seems therefore appropriate, particularly because for diversity-dependent or protracted single-birth-death models there is no standard species tree prior available which would have complicated our analyses. Furthermore, we do not expect that addition of diversity-dependence or protracted speciation alters the inference on whether speciation events take place simultaneously or not.

For our experiment, we used the default `pirouette` setup: when simulating a DNA alignment, the simplest (JC69) nucleotide substitution model and the simplest (strict) clock model are used. One could argue these models are overly simplistic and biologically irrelevant. We argue that this actually improves the clarity of our results for two reasons. The first reason is a practical one: because the computational load was lowered, we were

able to perform more replicates. For a process with this high amount of stochasticity, we think this is highly preferable. The second reason is that we think that this reduces the noise of our results, without hurting the experiment: in our setup, it is essential to assume the same site and clock model in inference as the one that is actually used in generating the data.

When simulating our phylogenies, we used an MBD model with and without extinction. In the case in which extinction is present, picking the BD tree prior as the generative tree prior seems justified: it is the best-fitting standard tree prior. For the MBD model without extinction, we could have picked the Yule tree prior. We believe, however, that this is of minor importance.

For the four MBD parameters λ, μ, ν and q, we investigated 1, 2, 4 and 3 different values respectively. We chose to use only one value for λ, because the proportion of multiple-birth events depends on the ratio between λ and a combination of ν and q. The crown age we selected for our experiments was based on trial and error of pilot experiments. Because the number of species in an MBD process is expected to increase exponentially, increasing the crown age has a profound effect on the number of taxa, with the consequence of having runs that took days to calculate. We decided to priortize the number of replicates at the cost of a lower number of taxa.

Ideally, we would have liked all our ESSes to be above the recommended value of 200. We are aware ESSes can be increased easily by making the MCMC chain longer. However, the bimodal distribution of ESSes is disadvantageous: to reduce the percentage of $ESS < 200$ by fifty percent, we would have had to increase the MCMC chain lengths by a factor of forty. We preferred to invest this run-time in doing more replicates.

This research measures the impact that the use of a non-MBD tree prior has on the inference error in phylogenetic construction for species that are subject to species pump dynamics. We found that the trees constructed with the standard BD species tree prior are similar (as measured by nLTT) to the true tree. This may be considered to be good news, as the implementation of an MBD species tree prior does not seem necessary. However, our results also suggest that it will be difficult to determine the parameter of the MBD process when one wants to fit the model to data of a system that is known to be subject to species pump dynamics. One way to test this would be to try birth-death skyline species tree priors and see if they pick up a signal of elevated speciation rates during co-occurring speciation events. These models may be prone to overparametrization because they have to assume a speciation rate for each interval in which the multiple speciation events take place, whereas the MBD model provides a dynamic explanation for these elevated speciation rates. But if they do pick up a signal, then it may still be worth implementing a species tree prior for the MBD model in Bayesian phylogenetic reconstruction tools.

REFERENCES

Bilderbeek, R.J.C. & Laudanno, G. (2019) *pirouette: create a posterior from a phylogeny*.

Bilderbeek, R.J.C., Laudanno, G. & Etienne, R.S. (2020) Quantifying the importance of an inference model in bayesian phylogenetics.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M.A., Ram-

baut, A. & Drummond, A.J. (2014) Beast 2: a software platform for bayesian evolutionary analysis. *PLoS computational biology*, **10**, e1003537.

Drummond, A.J. & Bouckaert, R.R. (2015) *Bayesian evolutionary analysis with BEAST*. Cambridge University Press.

Drummond, A.J. & Rambaut, A. (2007) Beast: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, **7**, 214.

Etienne, R.S., Haegeman, B., Stadler, T., Aze, T., Pearson, P.N., Purvis, A. & Phillimore, A.B. (2012) Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proc R Soc Lond B: Biol Sci*, **279**, 1300–1309.

Etienne, R.S., Morlon, H. & Lambert, A. (2014) Estimating the duration of speciation from phylogenies. *Evolution*, **68**, 2430–2440.

Haffer, J. (1969) Speciation in amazonian forest birds. *Science*, **165**, 131–137.

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

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

Janzen, T. & Etienne, R.S. (2017) Inferring the role of habitat dynamics in driving diversification: evidence for a species pump in lake tanganyika cichlids. *bioRxiv*.

Janzen, T., Höhna, S. & Etienne, R.S. (2015) Approximate bayesian computation of diversification rates from molecular phylogenies: introducing a new efficient summary statistic, the nltt. *Methods in Ecology and Evolution*, **6**, 566–575.

Laudanno, G. (2018) *MBD: Multiple Birth Death Diversification*. R package version 0.1.

Maddison, W.P., Midford, P.E. & Otto, S.P. (2007) Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, **56**, 701–710.

Muellner-Riehl, A.N., Schnitzler, J., Kissling, W.D., Mosbrugger, V., Rijsdijk, K.F., Seijmonsbergen, A.C., Versteegh, H. & Favre, A. (2019) Origins of global mountain plant biodiversity: Testing the 'mountain-geobiodiversity hypothesis'. *Journal of Biogeography*, **46**, 2826–2838.

Oaks, J.R., Siler, C.D. & Brown, R.M. (2019) The comparative biogeography of philippine geckos challenges predictions from a paradigm of climate-driven vicariant diversification across an island archipelago. *Evolution*, **73**, 1151–1167.

Ronquist, F. & Huelsenbeck, J.P. (2003) Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Sturmbauer, C., Baric, S., Salzburger, W., Rüber, L. & Verheyen, E. (2001) Lake level fluctuations synchronize genetic divergences of cichlid fishes in african lakes. *Molecular Biology and Evolution*, **18**, 144–154.

Valente, L.M., Phillimore, A.B. & Etienne, R.S. (2015) Equilibrium and non-equilibrium dynamics simultaneously operate in the galápagos islands. *Ecology Letters*, **18**, 844–852.

Verheyen, E., Rüber, L., Snoeks, J. & Meyer, A. (1996) Mitochondrial phylogeography of rock-dwelling cichlid fishes reveals evolutionary influence of historical lake level fluctuations of lake tanganyika, africa. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, **351**, 797–805.

5

SYNTHESIS

5.1. SUMMARY

This thesis can be summarized in one sentence: we've developed the tools to measure the error we make in phylogenetic inference and applied it on one non-standard speciation model.

The reason we did this, is because of our incomplete knowledge of speciation. Speciation generates biodiversity, which is important to us humans, through the ecosystem services provided by a species-rich environment. Understanding the evolutionary history, that is, knowing the phylogeny of multiple species, helps us understand the process of speciation. Because we do not know which macroevolutionary patterns (that is, again, a phylogeny) to expect exactly, we can only hope our estimated phylogenies are good enough. The work in this thesis helps to assess if our hopes are justified.

Within this chapter, I will put the work in this thesis into perspective. I will first take a look at the most basic thing produced, which is the software underlying the research, as this is the easiest to describe objectively. From this rather plain foundation, I will move on to the way the actual research is done and ending with the implications for the field of biology.

5.1.1. SOFTWARE

A simple way to quantify the amount of work is to count the lines of code and compare with related software. In figure 5.1 (and table 5.1) I show the number of (non-empty) lines of code for the packages I developed, the packages I maintain, the packages I contributed to, as well as BEAST2. BEAST2, which is the foundation of the work in this thesis, has the most lines of code, above 110k. After that comes `beautier` (27k lines), `phangorn` (18k), `pirouette` (17k), `daisieme` (14k), `DAISIE` (12k) and `razzo` (8k). `beautier` is an R package that creates a BEAST2 input file, and is part of the `babette` package suite, as described in chapter 2. `phangorn` is a general phylogenetics package of which I fixed some bugs. `pirouette` is the package described in chapter 3. `daisieme` is part of a project that did not reach full fruition yet (see below). `DAISIE` is an R package developed in our group, with 42 citations on Google scholar. `razzo` is the package described in chapter 4. Summing up the packages of which I wrote most of the code, results in 90k lines of code. This number of lines is still less than BEAST2 (with 110k), except all written in half the time. Also note that BEAST2 has 26 collaborators, of which 6 contributed more than 1k lines of code.

Quality Judging code by the number lines of code is simple, but this is irrelevant to estimate the quality of the software. Here I will highlight some indirect evidence of software quality, for the software listed in figure 5.1. To start with, all software in figure 5.1 uses a continuous integration (CI) service, which is known to significantly increase the number of bugs exposed (Vasilescu *et al.* 2015) and increases the speed at which new features are added (Vasilescu *et al.* 2015). A CI service is automatically activated when a developer puts a new version of his/her software online. The CI service will create a virtual computer from scratch, build the software and run it. These virtual computers can be of multiple operating systems. Where BEAST2 and `DAISIE` are tested on Linux

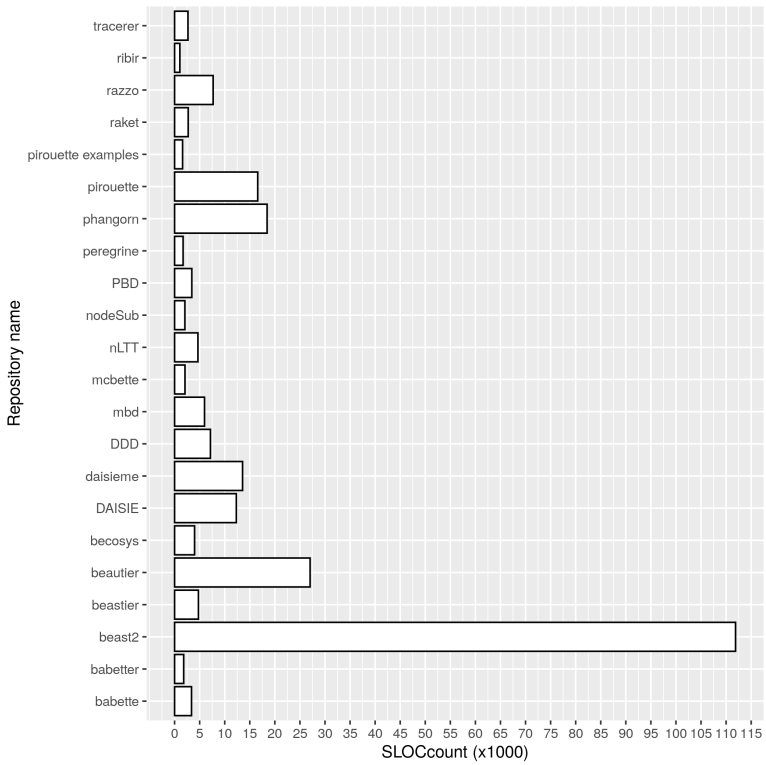


Figure 5.1 | SLOCcount: number of (non-empty) source code lines per repository

only, `beautier`, `pirouette` and my other R packages are tested to run under MacOS and Windows as well, assuring users of the three major operating systems can actually run these.

A simple metric to get an idea of code quality is the code coverage. Code coverage correlates with code quality (Horgan *et al.* 1994, Del Frate *et al.* 1995). The code coverage is the percentage of lines that is actually executed by tests. Writing tests is fundamental for writing quality code. These tests are usually run by the CI, each time a developer puts a new version online. Ideally all lines of code are tested. As can be seen in table 5.1, all `babette` packages have a 100% code coverage, compared to BEAST2, with an unknown/undisclosed code coverage, `phangorn` with approximately 70%, followed by `DAISIE` with approximately 60% [1].

Another measure to improve code quality is peer review. Similar to academic manuscripts, also code can be peer reviewed. For R code, rOpenSci is the non-profit organisation that does so. Note that a prerequisite for a code review by rOpenSci is that code coverage is 100%, therefore `phangorn` and `DAISIE` are not yet eligible. The five packages of the `babette` package suite have been reviewed, where `mcbette` is under review. The full process of the review of `babette` took approximately one year, as this is done in the free time of both me and the reviewers. Mostly due to this, there has not been time yet to have `pirouette` reviewed.

Relevance The relevance of software is another facet: one may write big pieces of software of high quality, but if nobody uses it, the work is still irrelevant.

One way to estimate the relevance is to measure the number of CRAN downloads per month. CRAN is a central repository for R packages, which keeps track of the number of downloads. By this measure, as of March 9th 20202, `phangorn` is most relevant, with 15k downloads per month, followed by `beautier` (975), `tracerer` (849) and `DAISIE` (736). Because BEAST2 is not an R package, it is absent from this list.

Another way to estimate the relevance is to measure the number of stars given on GitHub. GitHub is a website that hosts source code and that allows to develop software collaboratively. Logged-in users (there are 40 million) can give a star to a project to indicate his/her appreciation of the project. Going through the projects in 5.1, most stars are given, as of March 9th 2020, to BEAST2, with 134, followed by `phangorn` with 110 and `babette` with 20 stars. After `beautier` (6), `tracerer` (5), `beastier` (5) and `mcbette` (4), `pirouette`, `DAISIE` and `nLTT` have 3 stars. For repositories with 3 or less stars, these stars are given by the developers themselves and thus less relevant to indicate the relevance of a project.

Community A time-consuming aspects of developing software is taking care of its users, which includes the developer(s).

Users expect that R packages are easy to install. The R community has a centralized website from which packages can be installed easily, called CRAN (short for 'Comprehensive R Archive Network'). Therefore, a developer aims to get his/her package on CRAN. There are, however, many guidelines (see `https://cran.r-project.org/web/`

[1] I used the code coverage of the 'geodynamics' branch, as 'master' has 0 %

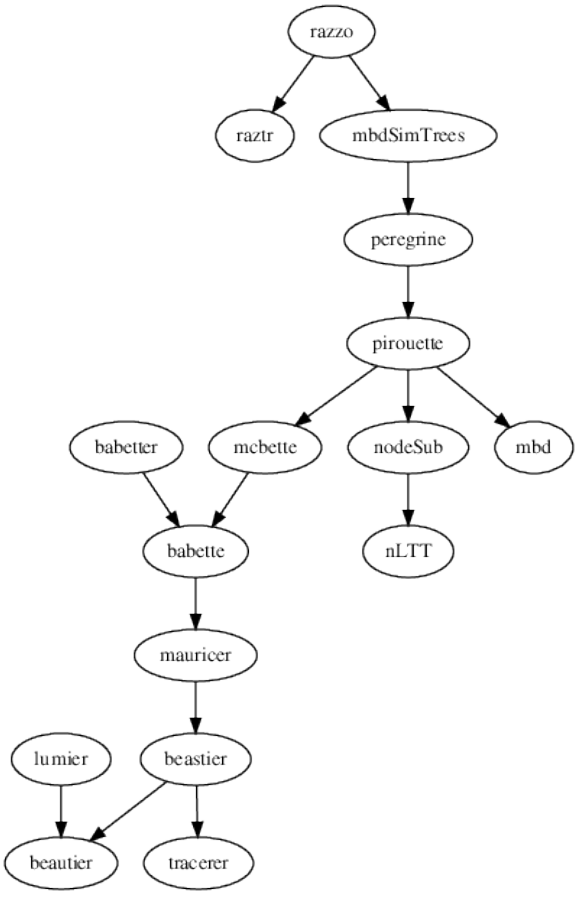


Figure 5.2 | Minimal spanning tree of the dependencies of the R packages used in this thesis. Arrows go from a package (at the tail), to the package it depends on (at the head). For example, 'beastier' depends on 'beautier'. The packages on CRAN are 'beautier', 'tracerer', 'beastier', 'mauricer' and 'babette'.

`packages/policies.html`) before a package gets accepted on CRAN. These guidelines exist to guarantee a minimum level of quality.

The most important guideline when submitting an R package to CRAN, is that all its dependencies are on CRAN. Figure 5.2 shows the dependencies of the R packages used in this thesis, showing that three out of the five `babette` packages depend on the two others. It would take one full year to get all packages on CRAN.

A consequence of taking care for the user, is that there should be a version of each of the packages that always works, regardless of ongoing development. If a top-level package, say `razzo` requires some different functionality of a bottom-level package such as `beautier`, there can be a cascade of new versions: a change in `beautier` can cause a change in any of the packages that depend on its. Due to this, `beautier` (as of 2020-03-10) is at its fourth CRAN version.

Users expect that the code they use has a certain quality, as they will depend on it. There are multiple ways to verify code quality. A popular feature is the use of status badges: dynamic images shown in the README of a project that signal a certain aspect of it, such as build status and code coverage. Additionally, code should be open, so the style and extent of tests can be verified. An example of a package that can improve in this regard is the `ape` package (Paradis and Schliep 2018), which contains a class for a phylogeny. It is possible to read the R code `ape` consists of from a CRAN submission. Except for that, there is no way to verify the code quality and development process: code is added by sending it per email, there is no website (such as GitHub) that tracks the development of the code and the tests are unavailable (although they apparently exist, according to personal communication with the maintainer of the package, Emmanuel Paradis).

Users also need documentation to learn to use a new package. One piece of documentation is an academic paper describing the functionality of a package. Such a paper is useful for getting the idea behind a package. User group meetings and tutorials (articles and videos) are better for learning how to use a package. BEAST2 has a user group meeting every half year, as well as dozens of tutorials (three of which I wrote). Specific to the R programming language is the vignette, a kind of documentation that can run a package's code. Counting the vignettes, `beautier` has four, `phangorn` has two, `pirouette` has six, `daisieme` has one (but well, it is unfinished), `DAISIE` has two and `razzo` has two. The complete `babette` package suite has 22 vignettes. Additionally, for `babette`, and `pirouette` there are nine video's to be downloaded or to be streamed from YouTube.

Users also expect a community: a place where they can ask questions, submit bug reports and contribute new code. GitHub has a checklist of seven recommended community standards (see, for example, `https://github.com/ropensci/babette/community`): having a one-line project description, having a README file, having a Code of Conduct, having a document that describes how to contribute, having specified a software license, as well as having template texts for Issues (among others, bug reports and feature request) and pull request (which is a code contribution of any type). Of these seven standards, BEAST2 has three, `beautier` all, `phangorn` has two, `pirouette` and `daisieme` have all, `DAISIE` has two and `razzo` has six.

5.1.2. SCIENTIFIC METHOD

Now that we have an idea of the amount of practical work underlying this thesis, let's take a look at the scientific methods used.

Reproduction in practice Reproducibility is an essential ingredient of science (McNutt 2014). The inability to reproduce experiments resulted in the so-called 'reproducibility crisis', which still is ongoing (Schooler 2014).

There are multiple threats to deliver reproducible science (Munafò *et al.* 2017). The two threats most relevant in the context of my research are HARKing and p-hacking (see figure 5.3 for all threats). HARKing, short for 'Hypothesis After Results are Known' is the practice to write down a hypothesis after having done an experiment. p-value hacking is the process of changing the analysis up until something significant is found.

The drawback of HARKing and p-hacking is that it leads to irreproducible science.

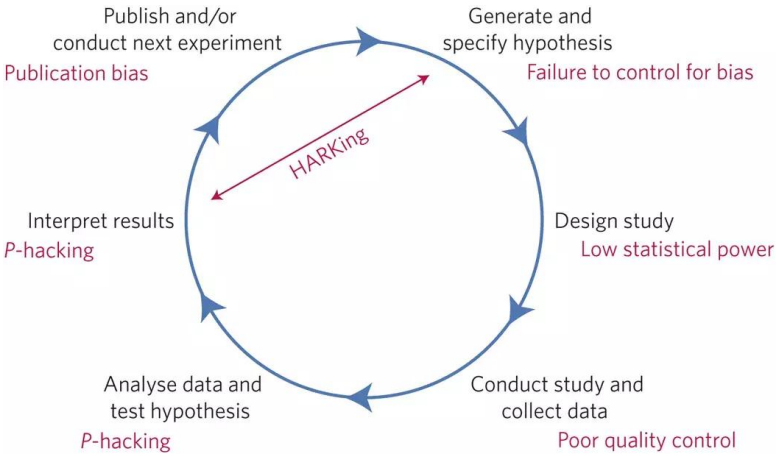


Figure 5.3 | Threats to reproducible science, from Munafò *et al.* 2017

From HARKing, hypotheses that were not under investigation, suddenly get some credibility, obtained from a random/no effect. p-value hacking gives more credibility to an experimental variable having an effect than warranted. It is estimated that 85% of the publications in biomedical sciences is a waste of resources (Chalmers and Glasziou 2009) (but note that this estimation is based on a logic reasoning, instead of empirical data), although the situation has improved since Macleod *et al.* 2014.

Assuring reproduction in practice One way to protect one's research from HARKing and p-hacking is the use of preregistration. Preregistration is the act of publishing an experiment's hypothesis, methods and analysis, before the experiment is finished.

Reproduction in this thesis The work in this thesis adheres to many of the best practices for reproducible research (Munafò *et al.* 2017). All papers in this thesis are Open Access. The `razzo` experiment was not pre-registered, as a lighter variant was used: code, manuscript and communication went via GitHub. GitHub is a website that allows people to collaborate. A feature of GitHub is that it keeps track of all changes. For `razzo`, the hypotheses and methods were written before the first results, and it is possible to verify this. Also the pilot runs of `razzo` can be found, as well as their results. By being completely open, we protected ourselves against HARKing and p-hacking.

Open Science Where reproducible research is an important facet of the scientific method, there is the Open Science movement that goes further: not only should there be openness in the research conducted, also the scientific article, resulting data, and software should be open. In that way, the scientific knowledge is accessible to all (among other, the tax payer) and can be reproduced by all.

All the academic articles have been put on bioRxiv before publication. bioRxiv is a pre-print server, meaning that it stores academic manuscripts before these appear in

print. Although the manuscript may not have been peer-reviewed, it is allowed to upload the version after peer-review, without the journal-specific layout. In that way, anyone can download my academic articles. Additionally, the GitHub repository that hosts the article is also accessible.

All the academic articles I published are Open Access. In this way, anyone can download them without any paywall.

All the academic articles are created by free and open source software ('FOSS'). Which means that anyone, regardless of operating system, can read these without any financial cost.

All the experiments are performed with FOSS only. Which means that anyone, regardless of operating system, can reproduce these, without any financial cost.

5.1.3. BIOLOGY

Now that we have an idea of practical work and scientific methods underlying this thesis, we can take a look what this thesis has contributed to increase our biological knowledge.

'babette' Because `babette` calls BEAST2, it is tempting to say that `babette` is just as relevant to the field of biology as BEAST2. This claim would be false, as not all aspects of BEAST2 are available within `babette`. The contribution of `babette` to the field of biology, is that it leads to more reproducible research: where it takes multiple programs to create, start and analyse a BEAST2 experiment, `babette` can do this from one R script.

'pirouette' The contribution of `pirouette` to the field of biology, is that it gives a thoroughly-tested framework to answer basic phylogenetic questions. The supplementary materials of `pirouette` shows plenty of examples that can evolve into a full academic paper when investigated more systematically.

Interestingly `pirouette` can also be used to investigate different models of how an alignment is simulated from a phylogeny, even though pirouette was not originally designed to do so. This was a fortunate example of the flexibility of `pirouette` and begs the question what `pirouette` will be most used for in the future.

'razzo' The contribution of `razzo` to the field of biology is the introduction of a new tree model, and measuring the error we make in our phylogenetic inference when nature follows a non-standard speciation model. This non-standard speciation model is the multiple-birth death (MBD) tree model, which is the first tree model that allows multiple speciation events to occur at exactly the same time. The predictions of `razzo` have always been straightforward: the stronger a tree violates the assumptions of a standard tree prior, the bigger the inference error made by that prior. What is unknown, is the extent to which this happens.

There are some assumptions that `razzo` makes that can be discussed, which are the assumptions of the MBD tree model and the assumptions made by the experimental setup. Where the MBD model assumes speciation events can co-occur at exactly the same time, one could easily argue that two speciation events at different locations cannot

happen at *exactly* the same time. The elegance of the MBD model is in the low number of parameters it needs to generate trees in which speciation can co-occur.

The biological relevance of this project hinges on multiple unknown facets. We did not investigate how common the MBD model is in nature, instead the model is loosely based on one example, which is the adaptive radiation in Lake Tanganyika. However, in the cases that MBD has a good fit with the data, the `razzo` experiment can show us the error we make in our phylogenetic inference. From this, we may either rest assured that our inference is good enough, or that we really need to add MBD to the set of standard models.

5.1.4. CANCELLED PROJECTS

During my thesis, I worked on some other projects that did not make it into this booklet. I will discuss these here.

'raket' `raket` is the ancestor of all chapters in this thesis. It would do the same thing as `razzo`, but for a different non-standard speciation model. This non-standard speciation model is called the Protracted Birth-Death model (PBD), in which speciation takes time: after a speciation event, one of the two new species is not directly recognized as such. Up until these are recognized, the number of species that are present (when looking back from the future) is underestimated.

The `raket` experiment would have one extra step compared to the `razzo` experiment: in the `raket` experiment, an *incipient* species tree would be simulated first, after which a species tree would be created from it. The way to do so, is by picking incipient species to represent a species.

One novel finding of the `raket` experiment, is that the sampling method to create a species tree is in some cases counterintuitive. The sampling method selects which incipient species will represent a (good) species. For example, one can select the incipient species that speciated most recently to represent its species. One would expect that sampling by this method would always result in a phylogeny that has the shortest branch lengths. This assumption is false, however, if there is a certain type of paraphyly, as shown in figure 5.4. As I was interested in obtaining phylogenies with the shortest branch lengths, the sampling method to obtain these was added. A similar story holds for sampling the oldest incipient species, which does not always result in a phylogeny with longest branches. Dueo to this, from the three existing sampling methods, two new methods have been added.

`raket` would be another illustration of the inference error we make if nature follows a non-standard speciation model. The same remarks as `razzo` apply here as well: it is unknown how well nature fits the PBD model. In the cases that nature fits the PBD model well, then `raket` would have been able to show the extent of the inference error.

'daisieme' `daisieme` (pronounce 'day-sham', similar to the French 'deuxième') is a project based on `DAISIE` (Etienne *et al.* 2019). `DAISIE` is an island model, which allows to estimate speciation, extinction and migration rates from one or more phylogenies. Island models, such as `DAISIE`, typically assume that the species on the mainland are fixed.

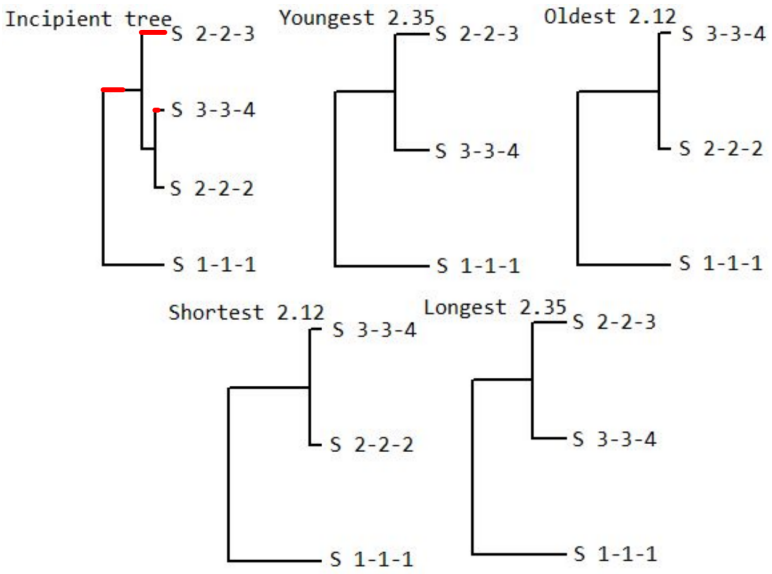


Figure 5.4 | Sampling of PBD trees. At the top-left is an incipient species trees, that shows four different incipient species. Species 'S2-2-2' and 'S2-2-3' are two different species, yet not recognized as such. The red edges denote a species still being an incipient species. The other four phylogenies are the result of four sampling methods. The sampling methods is shown above each phylogeny, as well as the sum of the branch lengths.

`daisieme` would investigate this assumption, by simulating phylogenies that do have mainland extinctions, estimating DAISIE parameters and comparing these parameters and predictions based on these parameters (such as the number of species and the number of colonizations) to the true parameter values and the true dynamics (of e.g. number of species and number of colonizations).

One of the predictions of `daisieme` is that the immigration rate will be overestimated when mainland extinction takes place, that is, species colonize an island earlier than actually true. This prediction is caused by the estimated colonization time of species that we do not know the actual colonization time of. For such a species, the colonization time is estimated with help from (part of) the DNA sequences taken from all extant species. For the island species of unknown immigration time, the closest mainland relative is chosen. Of these two species, the time of their speciation event is estimated. This time is used as the earliest time a colonization event could have taken place, which is the best estimate possible given the amount of information available. In figure 5.5 this is depicted as the vertical dashed line at the right. However, when the direct mainland ancestor has gone extinct, the *ancestor* of the extinct mainland species will be compared to the island species, resulting in an earlier estimated colonization time. In figure 5.5 this is depicted as the vertical dashed line at the left. If colonization times are estimated to happen earlier, migration rate should go up.

`daisieme` shows the extent of the inference error we make, for varying levels of mainland extinction. We can assume that the inference error increases for increasing

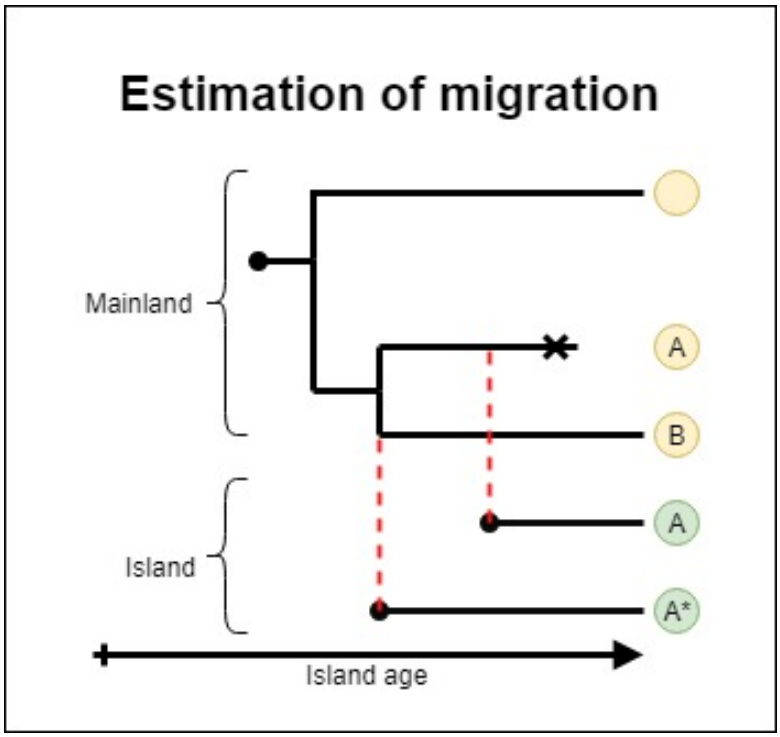


Figure 5.5 | daisieme example where we expect an overestimation in the migration rate. The top half shows the phylogeny of the mainland, which has tree species: A, B and an unlabelled one. The cross at the end of the yellow/mainland species A denotes its extinction. The vertical dashed red light directly left of it denotes the colonization of species A of the island, resulting in the green/island species A. The green/insland species A*, however, depicts the estimated colonization time, which is at the moment that mainland species A and B are formed.

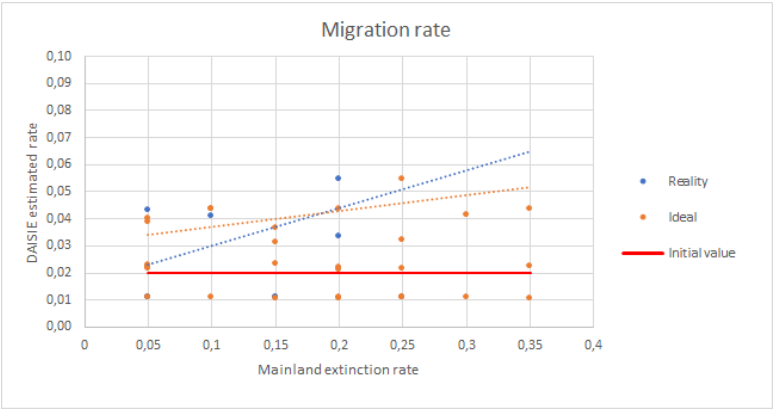


Figure 5.6 | Preliminary daisieme result: the estimated migration rate (vertical axis) for different mainland extinction rates (horizontal axis). Orange dots show the estimated mainland extinction rates in an ideal situation, which is when all immigration times are known. Blue dots show the estimated mainland extinction rates when the time of immigration needs to be estimated. The dotted lines show the result of a linear fit. The thick horizontal orange line shows the actual migration rate.

levels of mainland extinction, but the extent of this error will remain unknown for now.

One preliminary result, however, is shown in figure 5.6. This result points in the same way as the predictions, but a more thorough investigation is needed before drawing conclusions.

5.1.5. REFLECTION

When looking back at my PhD trajectory, I see some things that I will do again, things that I will avoid in the future, and some future work.

Things that I will do again It was inevitable that I would write exemplary software. Writing such software takes years, if not decades, of learning. Already a dozen of years before I started my PhD, I was reading the literature regarding software development. One could argue that, would I have done a worse job, I would have published more academic papers. I even agree on that! But for science as a whole, I think what I did is the superior way to go, where the cost of the few (that is, me) benefits the many. For me, it always hurts when some software developer does not care about his/her users, as I can easily envision the frustration this will cause.

Following the best practices for reproducible science is something I learned during my PhD trajectory and I will definitely continue (and improve) doing so. I think it was in my second year as a PhD student, when I noticed the question 'Do I believe this?' would pop up after a scientific talk. The answer, usually, was a no. First I thought that HARKing and p-hacking were even part of how science works and I did not want to become such a -in my eyes: fake- scientist. A presentation by Simine Vazire about Open Science showed me a way to conduct science in a way that would make me believe the result, among others to write an academic manuscript before having done the experiment. Since then,

I have taken that route. I am happy that if I ask myself 'Do I believe my own research findings?', that I can say yes.

Things that I will avoid Already early in my PhD work, the first ideas of `raket/razzo` were taking shape. Back then, I suggested not to pursue this line of research, because it would be clunky and inelegant. Clunky, because already one Bayesian phylogentic analysis takes hours. Unelegant, because of the backbone is just a factorial design of varying parameters. Nowadays, I still agree on this. I think, similar to other people in phylogenetics, that I should have pursued more light-weight and elegant experiments.

When developing a pipeline such as `pirouette`, there is a tension between (1) adding a new feature, (2) publish. The basic and minimal pipeline is the setup without candidate models and without twinning. Already this subset of the pipeline allows one to measure the inference error we make in phylogenetic inference. The `raket` paper, that was pre-registered two years ago, used only that part of the pipeline. Instead of investigating this minimal pipeline and publish the findings, features were added instead. These features are the use candidate models and the addition of a twin pipeline.

The first extra `pirouette` feature, which is the use of candidate models, has, in my opinion, caused mostly harm to the progress in my PhD work, without adding enough value. The main reason for this harm, is that the use of candidate models can only run under Linux and Mac, due to a feature of BEAST2 that only works under those two operating systems. Most desktop users, however, use the Windows operating system, so I needed to take this into account. Due to this, I had to write code that I would never run myself, a weird situation. Also, my co-author Giovanni Laudanno, who uses Windows, had a hard time to contribute to the `pirouette` code.

The second extra `pirouette` feature, which is the use of a twin pipeline, was, in my opinion, unwarranted to add before the publication of the minimal pipeline. There is some benefit to use twinning, but I think it would have been superior to show this benefit by reproducing an earlier `pirouette` publication with this new feature.

5.1.6. FUTURE WORK

My suggestions for future work are rather straightforward: (1) to measure the inference error we make on *standard* speciation models, (2) to measure the inference error we make on *other non-standard* speciation models, and (3) to make the MBD tree model part of the set of standard models.

Apply 'pirouette' on standard speciation models The goal of `pirouette` is the measure the inference error when a phylogeny is created by a non-standard tree model, but a standard tree prior is used in the inference. There are, however, only a couple of studies that investigate the inference error when using only standard tree models.

I think it would be useful to measure the inference error when using a standard tree model, when also assuming that tree model in the inference. This will give the baseline error, in a similar fashion as twinning does. This error would be the baseline error, in a similar way that twinning allows one to measure this. From these baseline errors, I would

enjoy to fit a mathematical model on these errors, to be able to obtain a prediction of this error without running the time-consuming Bayesian inference.

A next fundamental step would be to measure the inference error when creating phylogenies using a different standard tree model as is assumed in the inference. When we know the error we make when nature follows a BD model, when assuming a Yule model, this would give a sense of scale. It may even be that there is no reason to use BD at all, because the inference error is too little to warrant using it! Whatever this error, I would be curious to see how it compares to the errors found in `razzo`.

I understand why this has not been researched: it takes too long and there is little glory in finding this out. With `pirouette`, however, it should at least be easy to setup these experiments.

Apply 'pirouette' on multiple novel speciation models The inference error that `razzo` measures, is caused by the mismatch of using an MBD tree, yet assuming a BD tree model. It is easy to do this for any non-standard tree model, such as PBD, but also a time or diversity dependent tree model. Using different non-standard tree priors, gives us a better idea of when we can and when we cannot use our standard tree priors.

Add MBD tree prior to BEAST2 In `razzo`, we measure the inference error we make, when we generate and MBD tree and assume a BD tree model in our inference. What we do not know is the inference error would we assume an MBD tree model in our inference. Being able to use an MBD tree prior would give another baseline error: the inference error when nature follows MBD and we correctly assume this. To do so, the MBD tree prior must be added to BEAST2, `babette` should be able to use it, then a study similar to `razzo` can be done.

REFERENCES

B. Vasilescu, Y. Yu, H. Wang, P. Devanbu, and V. Filkov, *Quality and productivity outcomes relating to continuous integration in github,* in *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering* (ACM, 2015) pp. 805–816.

J. R. Horgan, S. London, and M. R. Lyu, *Achieving software quality with testing coverage measures,* Computer **27**, 60 (1994).

F. Del Frate, P. Garg, A. P. Mathur, and A. Pasquini, *On the correlation between code coverage and software reliability,* in *Software Reliability Engineering, 1995. Proceedings., Sixth International Symposium on* (IEEE, 1995) pp. 124–132.

E. Paradis and K. Schliep, *ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R,* Bioinformatics **35**, 526 (2018).

M. McNutt, *Reproducibility,* (2014).

J. W. Schooler, *Metascience could rescue the 'replication crisis',* Nature **515**, 9 (2014).

M. R. Munafò, B. A. Nosek, D. V. Bishop, K. S. Button, C. D. Chambers, N. P. Du Sert, U. Simonsohn, E.-J. Wagenmakers, J. J. Ware, and J. P. Ioannidis, *A manifesto for reproducible science,* Nature human behaviour **1**, 1 (2017).

I. Chalmers and P. Glasziou, *Avoidable waste in the production and reporting of research evidence,* The Lancet **374**, 86 (2009).

M. R. Macleod, S. Michie, I. Roberts, U. Dirnagl, I. Chalmers, J. P. Ioannidis, R. A.-S. Salman, A.-W. Chan, and P. Glasziou, *Biomedical research: increasing value, reducing waste,* The Lancet **383**, 101 (2014).

R. S. Etienne, L. M. Valente, and A. B. P. . B. Haegeman, *DAISIE: Dynamical Assembly of Islands by Speciation, Immigration and Extinction* (2019), r package version 1.6.1.

5.2. SUPPLEMENTARY MATERIALS

In these supplementary materials, I show the raw data referred to in the main text.

5.2.1. ALTMETRICS

- 60,000 GitHub commits, 1.1k repositories, 421 stars, 242 followers
- 133 YouTube videos, 68 subscribers, 11k views
- Supervised 2 MSc students
- Supervised 6 BSc students
- Supervised 7 interns from secondary schools
- Organised 172 social events
- Since Jan 2017, presented 20 times at TECE
- Publish 5 packages on CRAN
- Passed rOpenSci peer-review for 4 R packages
- Taught +220 evenings about programming

name	title	sloccount	cc	ns	ndm	ndt
aureole	R interface to the Encyclopedia of Life	460	100	0		
babette	Control 'BEAST2'	3378	100	20	452	1173
babette examples	All babette examples	149				
babetter	Check babette	1816		0		
beast2	BEAST2	111886		134		
beastier	Call 'BEAST2'	4757	100	5	709	4579
beautier	'BEAUti' from R	27030	100	6	975	7135
becosys	Unified Interface To Phylogenetics Models Of Speciation	3989	78	0		
DAISIE	Dynamical Assembly of Islands by Speciation, Immigration and Extinction	12314	0	3	736	18000
daisieme	Island Diversification With Mainland Extinction	13558	97	1		
DDD	Diversity-Dependent Diversification	7151	24	1	1951	70000
mauricer	Install 'BEAST2' Packages	519	100	1	742	2202
mbd	Multiple Birth Death Diversification	5972		1		
mcbette	Model Comparison Using 'babette'	2074	100	4		
nLTT	Calculate the NLTT Statistic	4658	99	3	702	23000
nodeSub	Simulate Sequences	2055	53	1		
PBD	Protracted Birth-Death Model of Diversification	3437	57	1	649	28000
peregrine	Work With The Groninger Peregrine Computer Cluster	1699	98	2		
phangorn	Phylogenetic Reconstruction and Analysis	18453	69	110	15000	420000
pirouette	Create a Bayesian Posterior From a Phylogeny	16584	99	3		
pirouette examples	All pirouette examples	1596				
raket	What If Speciation Takes Time?	2716	58	0		
raztr	Razzo Test Results	52		0		
razzo	The Error if Nature is MBD	7690	76	2		
ribir	ribir, basic phylogenetics page	1053	95	0		
tracerer	Tracer from R	2671	100	5	849	5359

Table 5.1 | Repository features. name: the CRAN package name. title: the R package title, as taken from the DESCRIPTION filesloccount: the number of (non-empty) lines of code.cc: code coverage, as a percentage, where 100 percent denotes that all code is covered by tests. ns: number of stars on GitHub. ndm: number of CRAN downloads per monthndt: total number of CRAN downloads

SUMMARY

THIS summary is written especially for non-biologists, so they can understand what is discussed in this thesis.

Speciation There are plenty of (animal, plant, etc.) species on the world. In Earth's early days this was not yet the the case: it took hundreds of millions of years for the first species to arise. In the many years that followed, billions of species have formed. The process that creates new species, we call speciation.

Speciation in bacteria There are multiple ways that speciation can occur. For bacteria, we state that two bacteria are of different species, if their DNA differs enough. Bacteria multiply when the environment is suitable and with each cell division, the DNA of the new bacteria changes slightly. Thus, when one starts with two identical bacteria, after some time, one ends up with two different bacterial species.

Speciation in animals For animals it is harder to state when two animals are of two different species. A commonly used definition is that two groups of animals are of different species, when the offspring of individuals of the different groups is either absent or results in infertile grandchildren.

There are multiple mechanisms that cause speciation in animals. One simple mechanisms is the split of group in two groups by a change in the environment, as can be done by a river or a mountain.

Phylogenies When we look at multiple species over a longer period of time (that is, millions of years!), likely there will be speciation events. Some species will give rise to more new species than others. We can display this process using a phylogenetic tree, as shown in figure 5.7.

DNA One needs some form of information to base a phylogeny on, such as, for example, the DNA sequence of the species within the phylogeny. All living beings have DNA, thanks to which it is possible to put all species in one big phylogeny. Each time DNA is transferred to a next generation, it changes a little bit. Due to this property, it is possible to base a phylogeny on DNA sequences. Simply put: species that have a more similar DNA sequence, are closer related.

Phylogenetic model There are multiple ways to construct a phylogeny from DNA sequences, because one can have different assumptions regarding how speciation occurs. For example, one can assume that speciation events occur just as often all the time (on

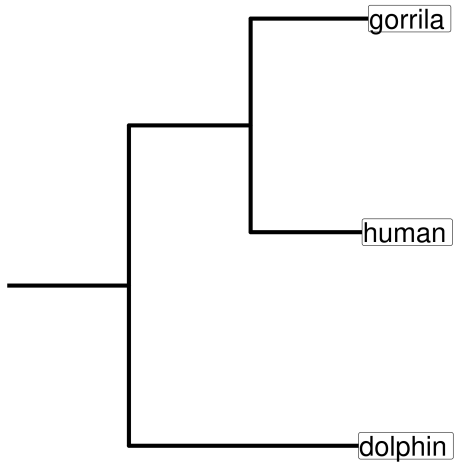


Figure 5.7 | A phylogeny that displays that human's and gorilla's are more related to one another, compared to humans and dolphins. This phylogeny is not to scale.

average) in all species. Or one may assume that DNA of all species have the same mutation rate. The collection of assumptions, we call a model. In this case, we call it a phylogenetic model.

Constructing phylogenies There are computer programs that construct a phylogeny from a phylogenetic model and DNA sequences. One of the most popular of such programs is called BEAST2. Because I would simulate many thousands of phylogenies, I needed to be able to do so from scripts only (that is, without any mouse clicks). For that reasons, I programmed `babette`, an R package with which one can call BEAST2. In chapter 2, one can read more about `babette`.

Phylogenetic models Because one can pick many different assumption regarding speciation, the question which one is best arises quickly. And there are also multiple methods to compare phylogenetic models (that is, to find out which one is 'best'). A drawback of most methods is that the phylogenetic model needs to be understood well mathematically. This means, that before one can measure how 'good' a new phylogenetic model is, it needs to be solved mathematically first.

Determine how good phylogenetic models are Giovanni, Rampal and I invented a way to determine if it is important to solve a new phylogenetic model analytically. With our new method, one only needs to simulate a lot of phylogenies using the novel method. Usually, this is way easier than solving a model mathematically. This method was put in an R package called `pirouette`. In chapter 3 one can read more about `pirouette`.

Testing a new speciation model After we invented a method to determine how important it is to solve a phylogenetic model mathematically, we applied the method on a new

phylogenetic model that has not yet been solved mathematically.

This new model is named the MBD ('Multiple-Birth Death') model and was invented by Giovanni. In this model, one assumes that speciation occurs in all species equally often on average, except that sometimes a 'speciation wave' occurs, in which multiple species speciate at the same time. In chapter 4 one can read how we did this exactly.

We found out, that would nature follow the MBD model, we can make phylogenies using simpler models that are just as good.

Conclusion This thesis shows that we can find out whether or not we should investigate a novel phylogenetic model in-dept, so that scientists can better spend their time.

The nice thing about my research is that other scientists also profit from it: with `babette` anyone can easily constrict phylogenies from DNA sequences. At the moment of writing, there have been 3 scientific publications that use `babette`. Also `pirouette` has become a strong R package, but without any citations yet.

SAMENVATTING

DEZE Nederlandse samenvatting is speciaal geschreven voor niet-biologen, zodat zij een beter idee kunnen krijgen wat er in dit proefschrift besproken wordt.

Soortvorming Er zijn op de wereld veel verschillende (dier-, plant-, etc.) soorten. Helemaal in het begin van het ontstaan van de Aarde, was dit nog niet zo, want toen ontstonden de eerste soorten. In de loop van de tijd zijn er veel soorten bijgekomen. Het proces dat hiervoor zorgt, noemen we soortvorming.

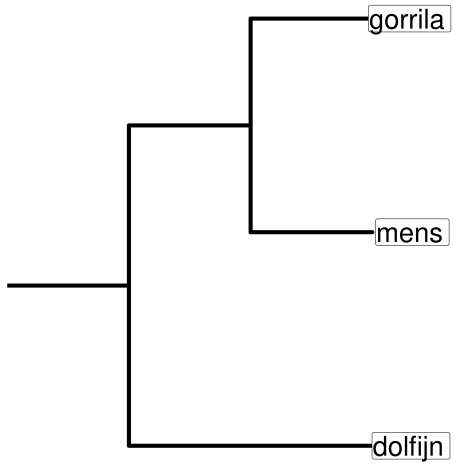
Soortvorming in bacteriën Soortvorming kan op meerdere manieren gebeuren. In bacterieën zeggen we dat twee bacteriën verschillend zijn, als hun DNA genoeg verschilt. Bacteriën vermenigvuldigen zichzelf als de omstandigheden gunstig zijn en bij elke celding vinden er veranderingen in het DNA plaats. Als je twee identieke bacteriën lang genoeg laat delen, heb je dus na een tijd twee verschillende bacteriesoorten.

Soortvorming in (vooral) dieren Bij dieren is het moeilijker te zeggen wanneer twee dieren verschillende soorten zijn. Een veelgebruikte definitie is dat twee groepen dieren verschillende diersoorten zijn, als een kruising tussen de twee groepen geen of onvruchtbare kleinkinderen oplevert.

Er zijn meerdere mechanismen die ervoor zorgen dat soortvorming in dieren plaatsvind. Een simpel mechanisme is dat een groep dieren een tweeën gesplits wordt door een verandering in het landschap, zoals een rivier of bergketen.

Fylogenieën Als we kijken naar een verzameling soorten over langere tijd (denk aan miljoenen jaren!), dan zal er waarschijnlijk soortvorming plaatsvinden. Sommige soorten zullen meer nieuwe soorten dan anderen opleveren. We kunnen dit proces laten zien met een fylogenetische boom, zoals bijvoorbeeld in figuur 5.8.

DNA Er is informatie nodig om een fylogenie te kunnen maken, bijvoorbeeld de DNA volgorde van de soorten in de fylogenie. Alle levende wezens hebben DNA, waardoor het mogelijk is alle soorten in een grote fylogenie te zetten. Elke keer dat DNA wordt doorgegeven aan de volgende generatie, verandert de DNA volgorde een klein beetje. Deze eigenschap maakt het mogelijk om een fylogenie te kunnen baseren op DNA volgordes. Simpel gezegt: de soorten waarvan de DNA volgordes het meest op elkaar lijken, zijn meer aan elkaar verwant.



Figuur 5.8 | Een fylogenie die laat zien dat mensen en gorilla's meer aan elkaar verwant zijn dan mensen en dolfijnen. Deze fylogenie is niet op schaal.

Fylogenetisch model Er zijn meerdere manieren om een fylogenie te maken aan de hand van DNA volgordes, omdat je verschillende aannames kunt hebben over hoe soortvorming plaatsvind. Je kunt bijvoorbeeld aannemen dat soortvorming gemiddeld altijd even vaak optreedt. Of dat DNA in alle soorten altijd even snel verandert. De verzameling van aannames noemen we een model, in dit geval noemen we dit een fylogenetisch model.

Fylogenieën maken Er zijn computerprogramma's die van een fylogenetisch model en DNA volgorden een fylogenie kunnen maken. Eén van de populairste is het programma BEAST2. Omdat ik veel fylogenieën zou gaan maken, was het voor mij belangrijk dat ik dit met enkel code (dus zonder muisklikken) zou kunnen doen. Daarom heb ik `babette` geprogrammeerd, een R package waarmee je BEAST2 kunt aanroepen. In hoofdstuk 2 kun je lezen over `babette`.

Fylogenetische modellen Omdat je veel aannames kunt maken over hoe soortvorming plaatsvindt, is het de vraag welke de beste verzameling aannames is. En ook het vergelijken van fylogenetische modellen (om uit te vinden welke 'de beste' is) kan op meerdere manieren. Een nadeel van de meeste manieren is dat het fylogenetische model wiskundig goed onderzocht moet zijn. Dit betekent dat je eerst een nieuw fylogenetisch model wiskundig moet oplossen, voor je kunt weten hoe goed dat model is.

Kijken hoe goed fylogenetische modellen zijn Ik, Giovani en Rampal hebben een manier bedacht om te kijken of het wel belangrijk is om een nieuw fylogenetisch model wiskundig op te lossen. Met onze manier hoef je alleen maar een boel fylogenieën van het nieuw model te simuleren. Dit is vaak veel gemakkelijker dan een model wiskundig op

te lossen. Deze manier hebben we in een R package gestopt, die we `pirouette` hebben genoemd. In hoofdstuk 3 kun je lezen over `pirouette`.

Een nieuw soortvormingsmodel testen Toen we een manier hadden om te kijken hoe belangrijk het is om een fylogenetisch model wiskundig op te lossen, gingen we dit gebruiken op een nieuw fylogenetisch model, dat nog niet wiskundig is opgelost.

Dit nieuwe model heet het MBD ('Multiple-Birth Death') model en is bedacht door Giovanni. Binnen dit model is de aanname dat soortvorming altijd voor alle soorten even vaak voorkomt, maar dat er soms een 'geboortegolf' optreedt, waarin er in meerdere soorten tegelijk soortvorming optreedt.

In hoofdstukken 4 kun je lezen hoe we dit precies hebben gedaan. We kwamen erachter dat zóu de natuur dit nieuwe en ingewikkeldere model volgen, je met bestaande en simpelere modellen net zo goede fylogenieen kunt maken. Dus als je alleen maar goeie fylogenieën wilt krijgen, is het niet nodig het MBD model op te lossen.

Conclusie Dit proefschrift leert ons dat we kunnen weten of we een nieuw fylogenetisch model zouden moeten onderzoeken, waardoor wetenschappers nuttigere dingen kunnen doen.

Het mooie aan mijn onderzoek is dat andere wetenschappers er zelf gemakkelijk ook wat mee kunnen: met `babette` kan iedereen gemakkelijk fylogenieen maken uit DNA sequenties. Op het moment van schrijven zijn er 3 wetenschappelijk artikelen gepubliceerd die `babette` gebruiken. Ook `pirouette` is een sterk R package geworden, maar er zijn nog geen publicaties die het gebruiken.

CURRICULUM VITÆ

Richèl J.C. Bilderbeek

02-09-1980	Born in Milsbeek, The Netherlands.

EDUCATION

1999–2005	Undergraduate in Biology Rijksuniversiteit Groningen
2007–2008	Undergraduate in Pre-higher Education in Biology Rijksuniversiteit Groningen
2011–2014	Undergraduate in Mechatronics Alfa College Groningen
2014-2019	PhD. Theoretical Biology Rijksuniversiteit Groningen *Thesis:* Speciation and the error we make in phylogenetic inference *Promotor:* Prof. dr. R. S. Etienne

WORK EXPERIENCE

2004–2018	Software developer, many projects
2008–2010	Secondary school teacher, in stagecraft, biology, physics, chemistry VMBO 't Venster Arnhem

VOLUNTEERING

2014–now	Coordinator of courses in programming and digital electronics Stichting De Jonge Onderzoekers Groningen
2000–2012	Light and sound technician Prinsentheater Groningen