Phylogenetics

Introduction to BEAST2

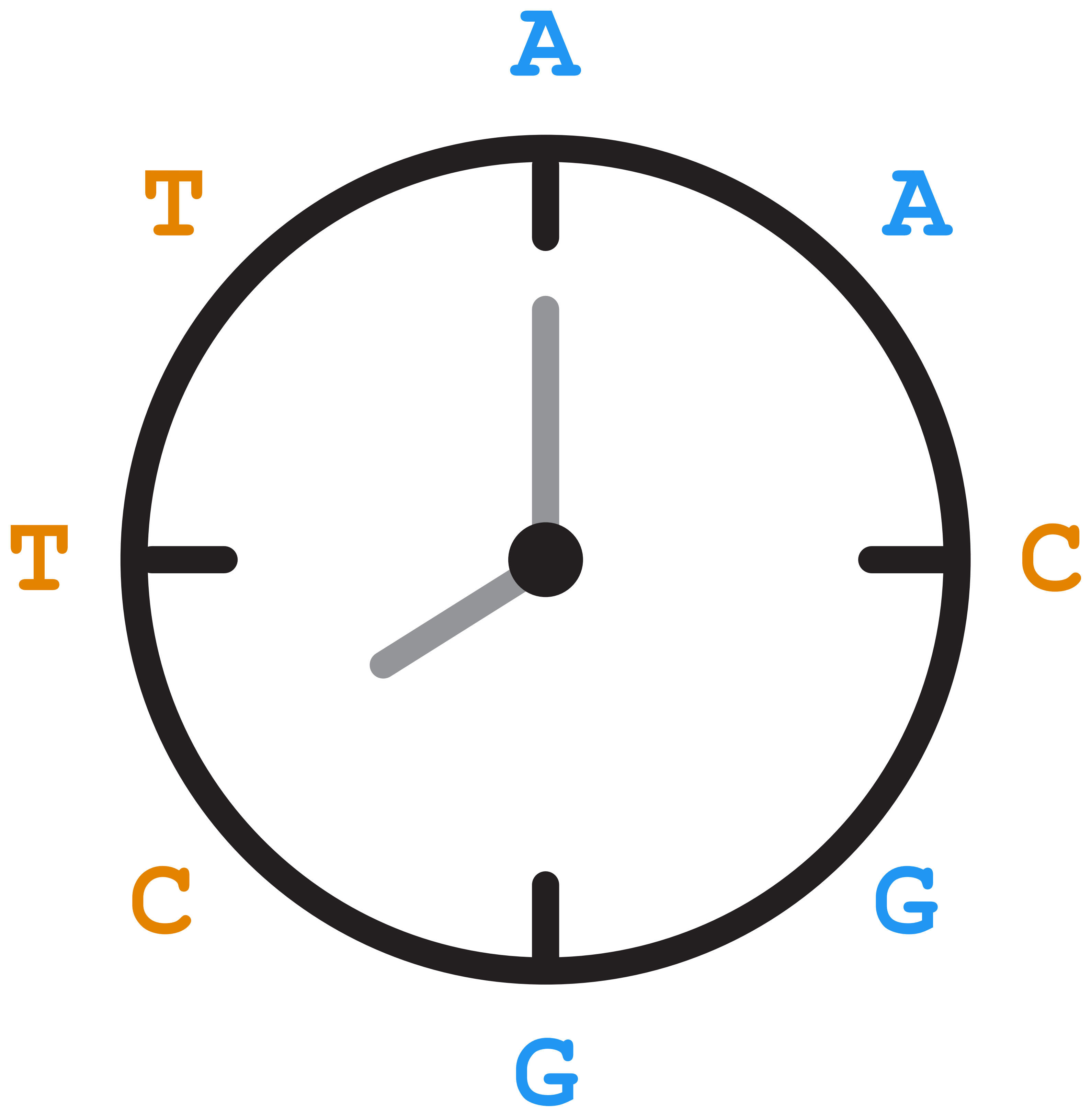
RL-V3 MPP

Rachel Warnock

17.06.25

Today's objectives

- Intro to BEAST2
- Node dating using BEAST2



Recap

How do we find the ‘best’ tree?

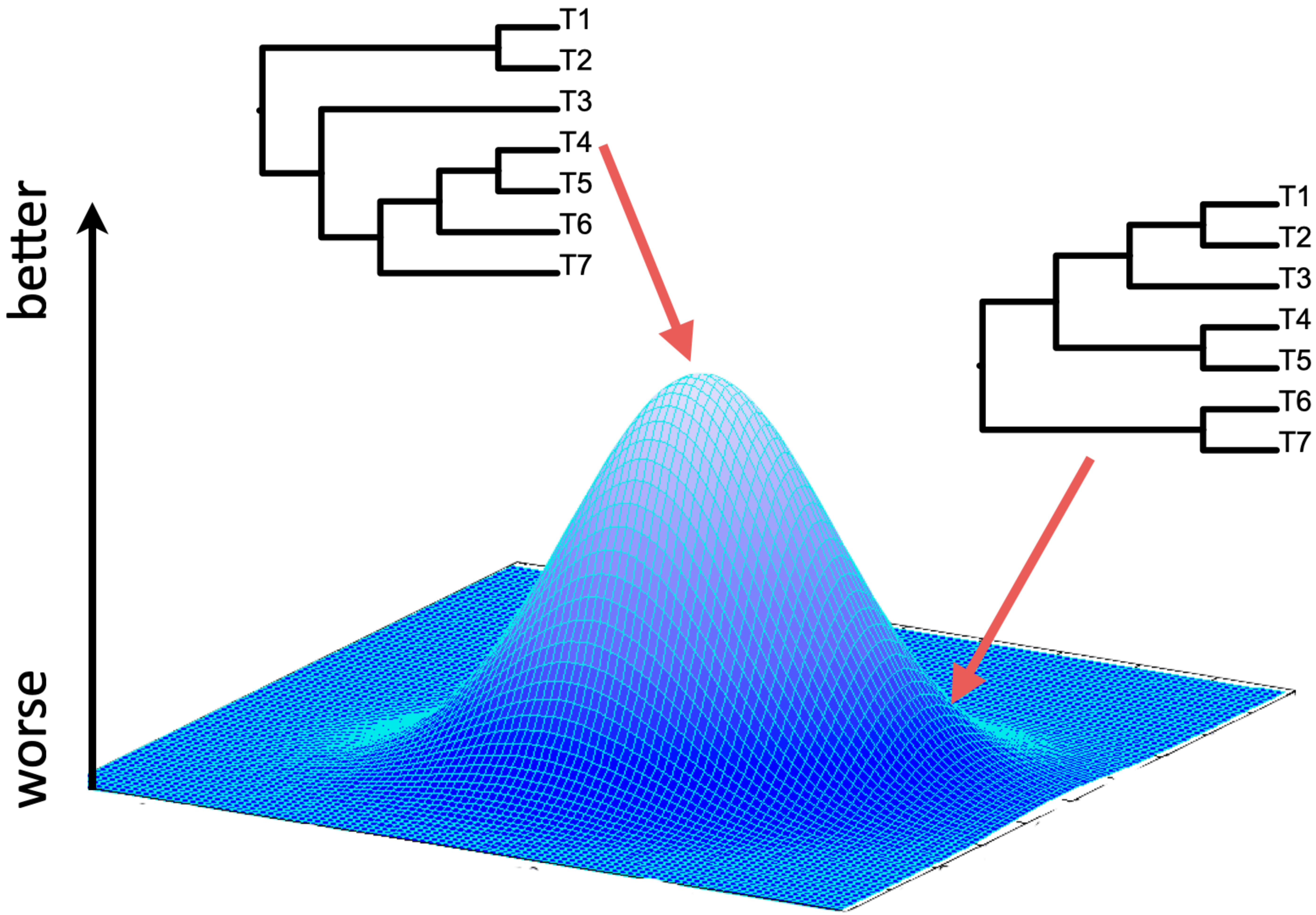


Image source Tracy Heath

It depends how you measure ‘best’

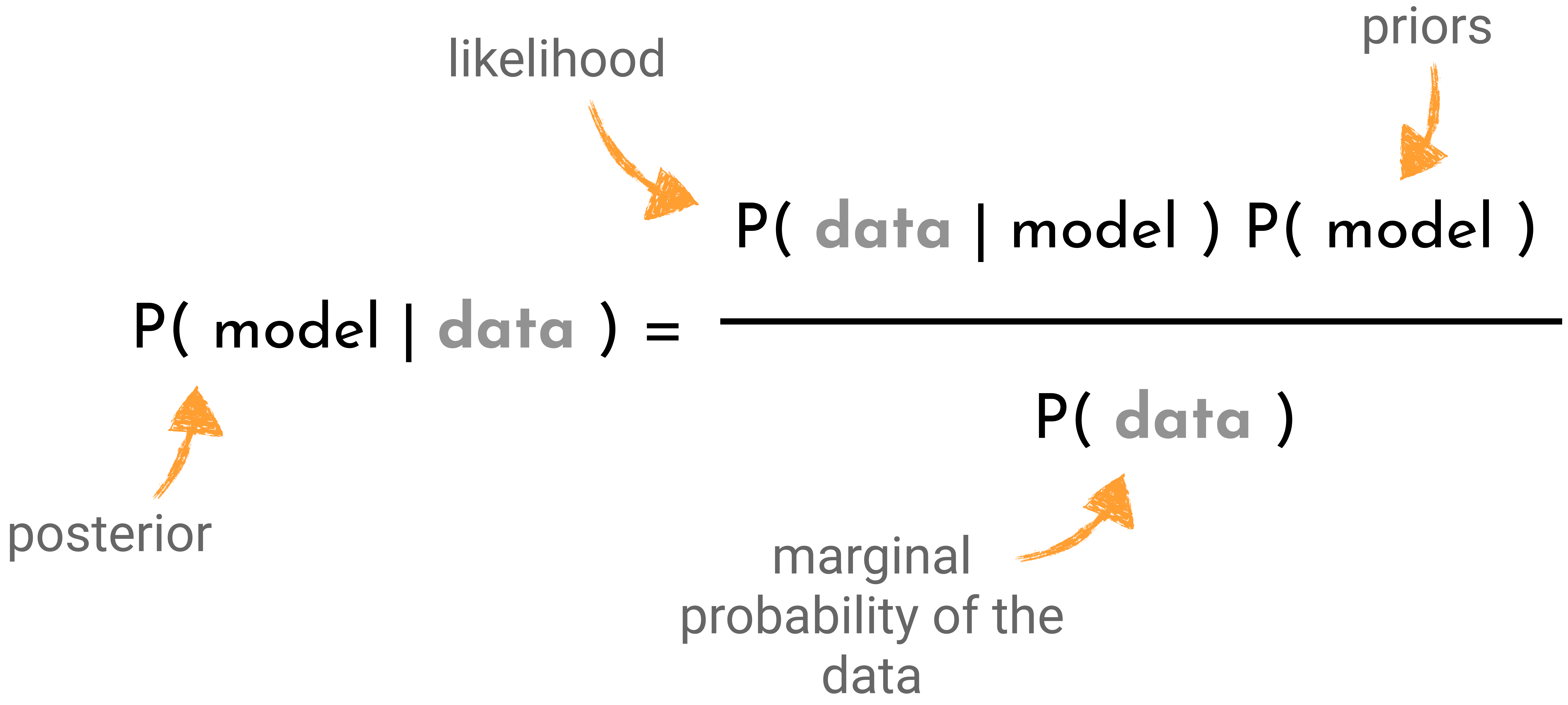
Method	Criterion (tree score)
Maximum parsimony	Minimum number of changes
Maximum likelihood	Likelihood score (probability), optimised over branch lengths and model parameters
Bayesian inference	Posterior probability, integrating over branch lengths and model parameters

Both maximum likelihood and Bayesian inference are model-based approaches

Note these are not the only approaches to tree-building but they are the most widely used

Bayesian divergence time estimation

We use a Bayesian framework



likelihood

priors

$$P(\text{ model } | \text{ data }) = \frac{P(\text{ data } | \text{ model })\, P(\text{ model })}{P(\text{ data })}$$

posterior

marginal probability of the data

Bayesian divergence time estimation

The data

and / or

0101... ATTG...
1101... TTGC...
0100... ATTC...

phylogenetics characters

sample ages

Understanding the tripartite approach to Bayesian divergence time estimation
Warnock, Wright. (2020)

Bayesian divergence time estimation

The data

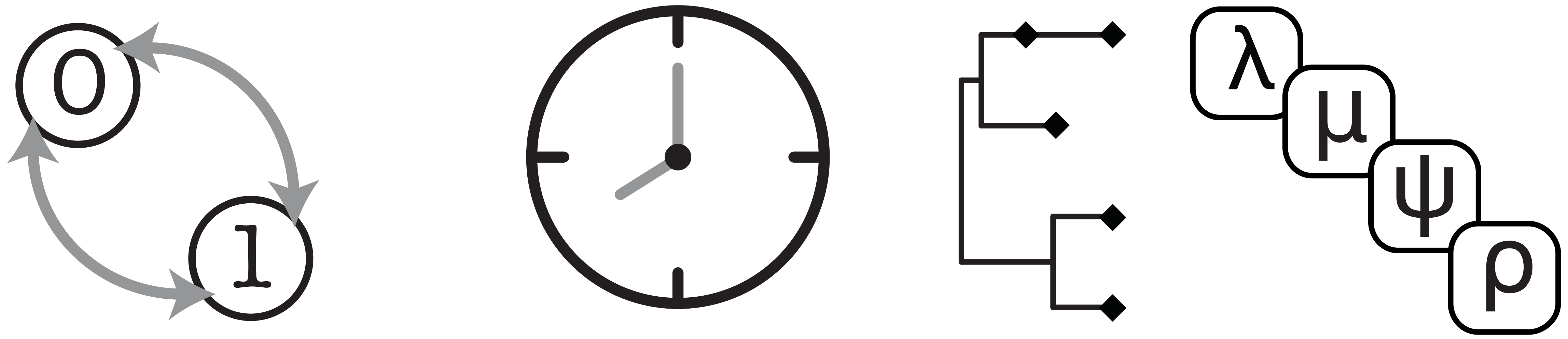
and / or

0101... ATTG...
1101... TTGC...
0100... ATTC...

phylogenetics characters

sample ages

3 model components



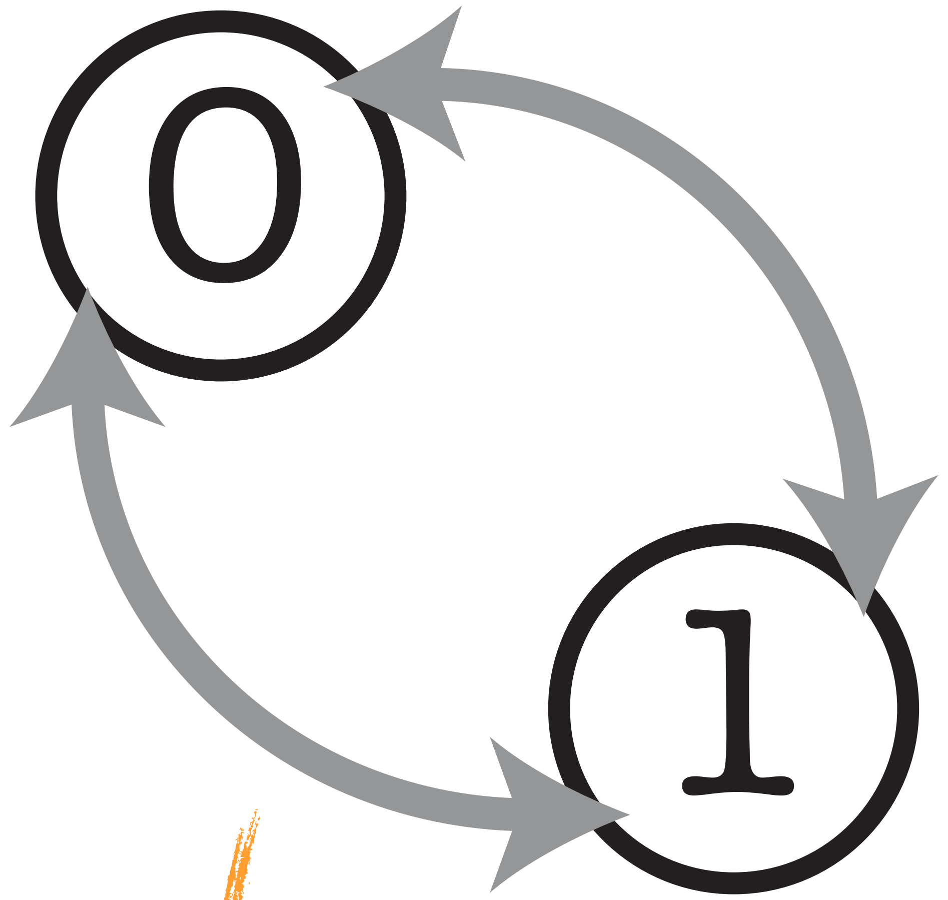
substitution model

clock model

tree and tree model

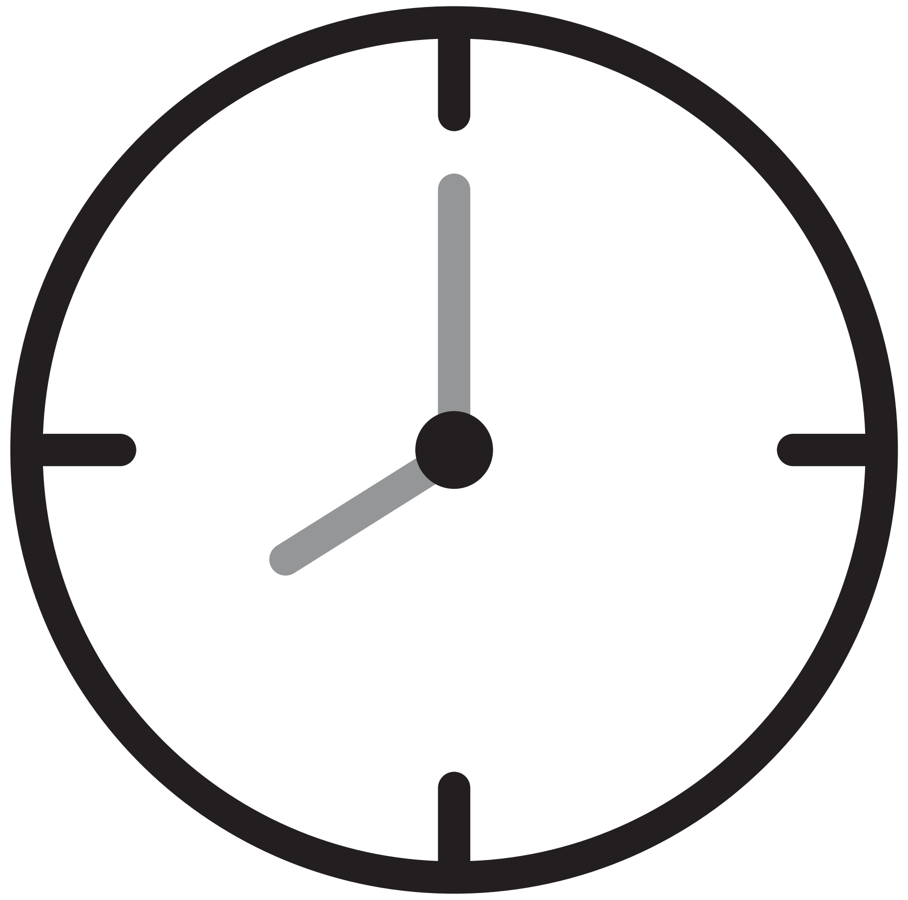
Understanding the tripartite approach to Bayesian divergence time estimation
Warnock, Wright. (2020)

substitution model

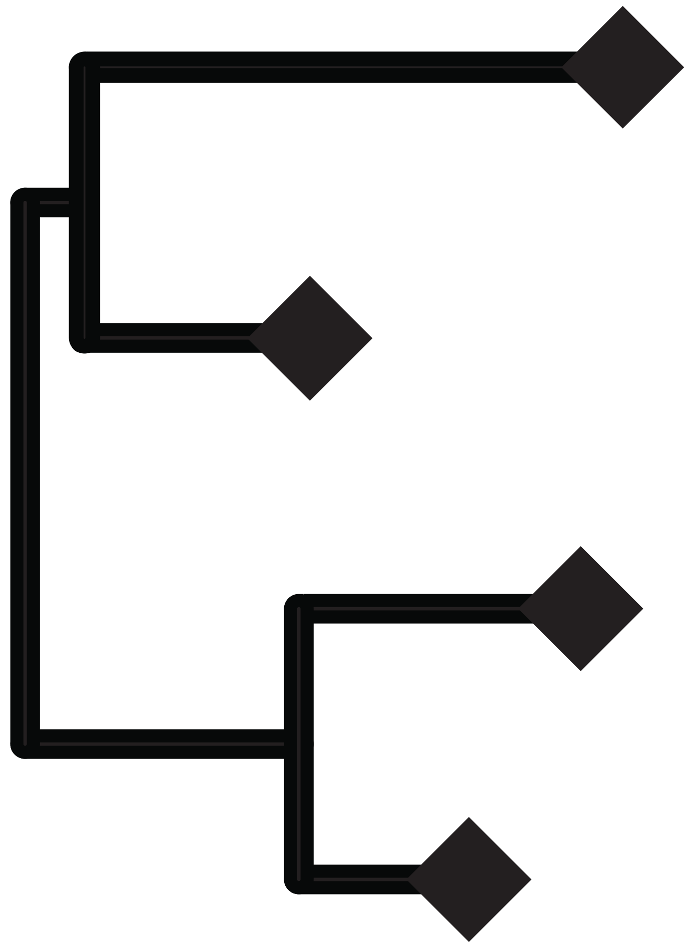


How likely are we to observe a change between character states? e.g., A → T

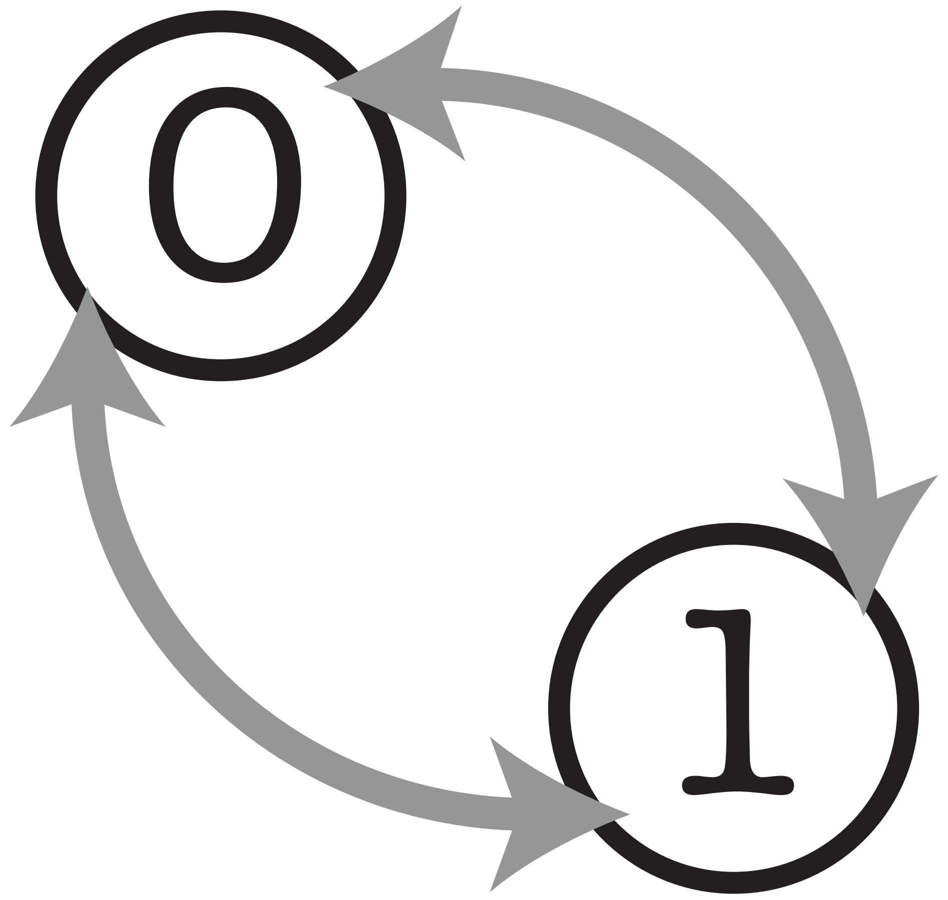
clock model



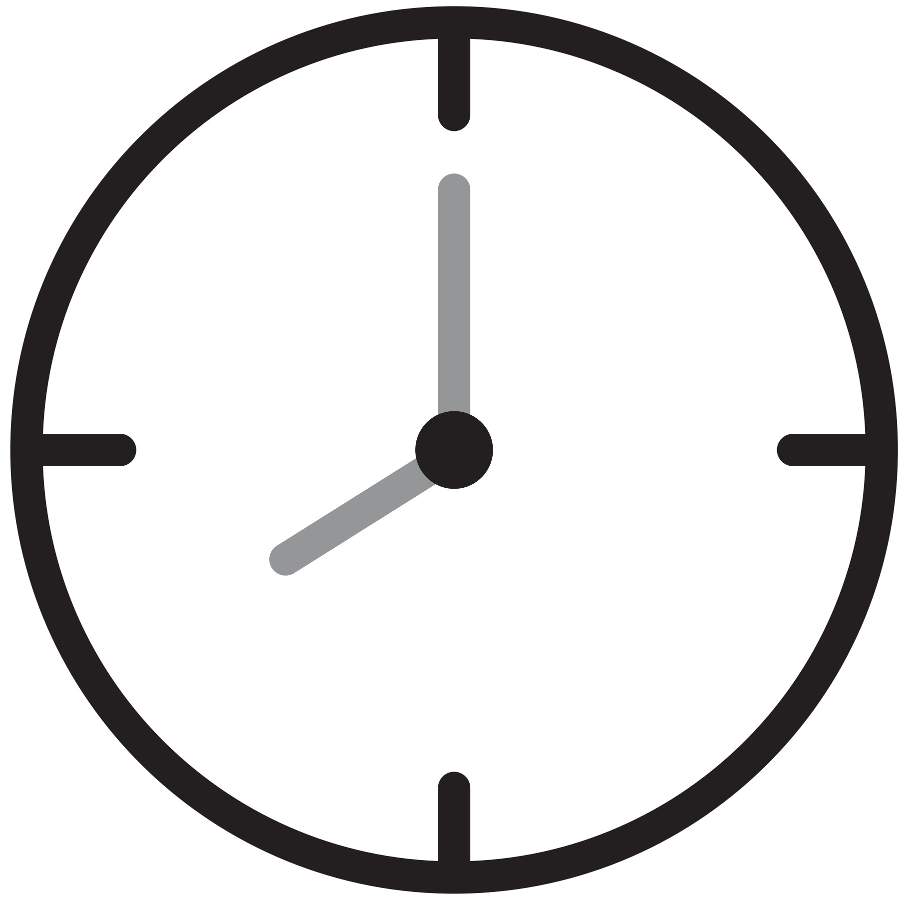
tree model



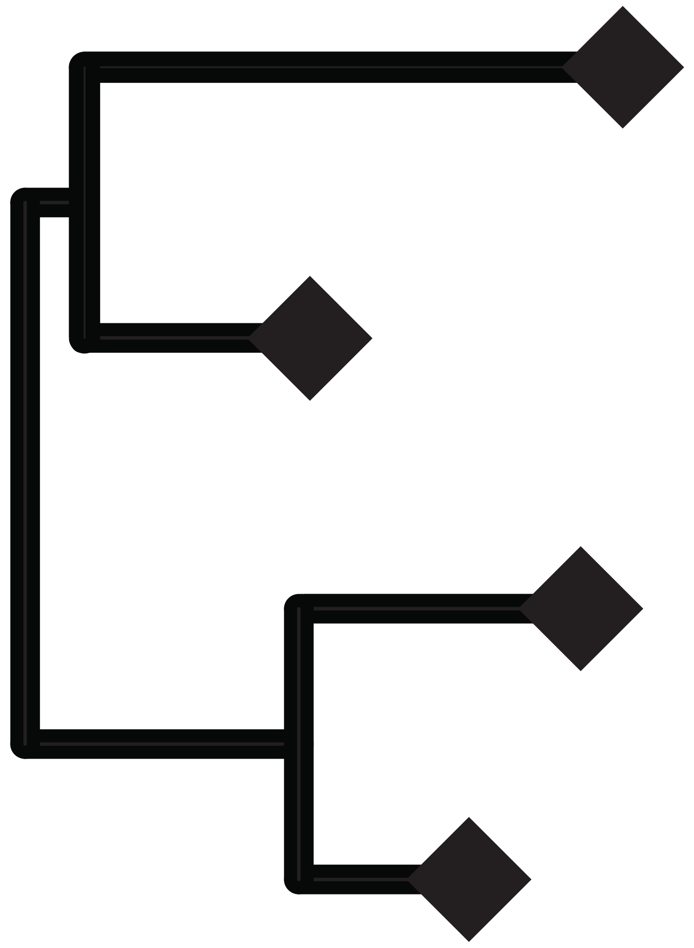
substitution model



clock model

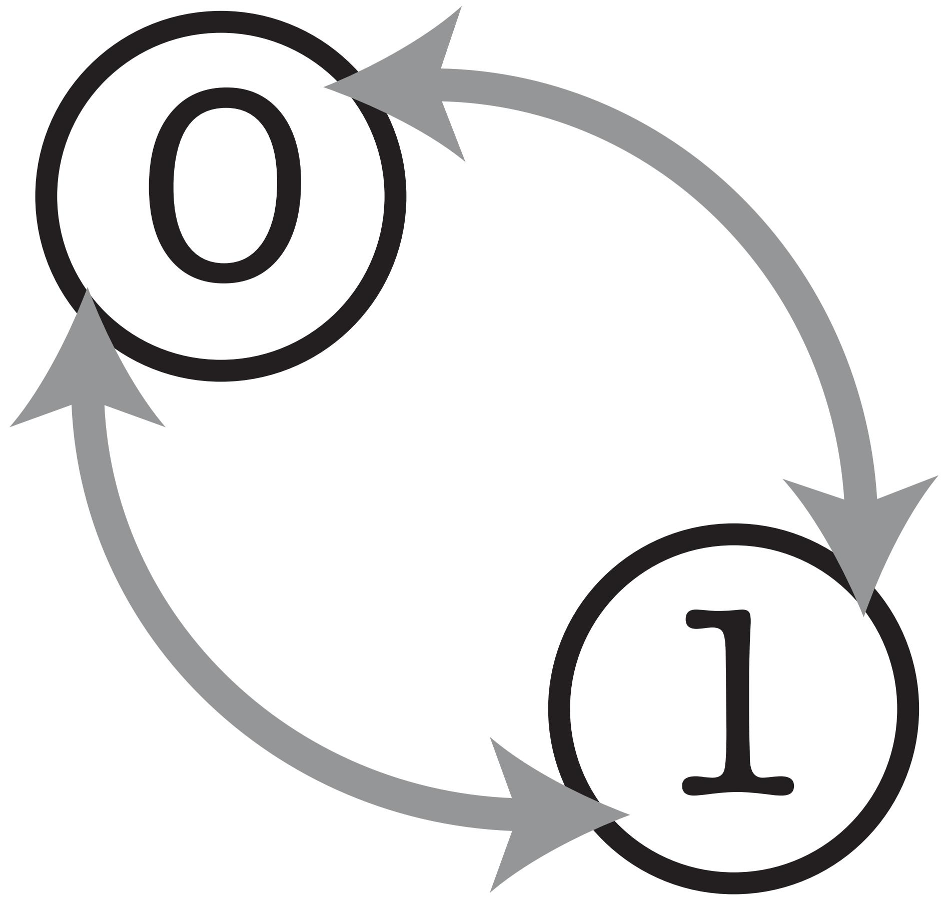


tree model

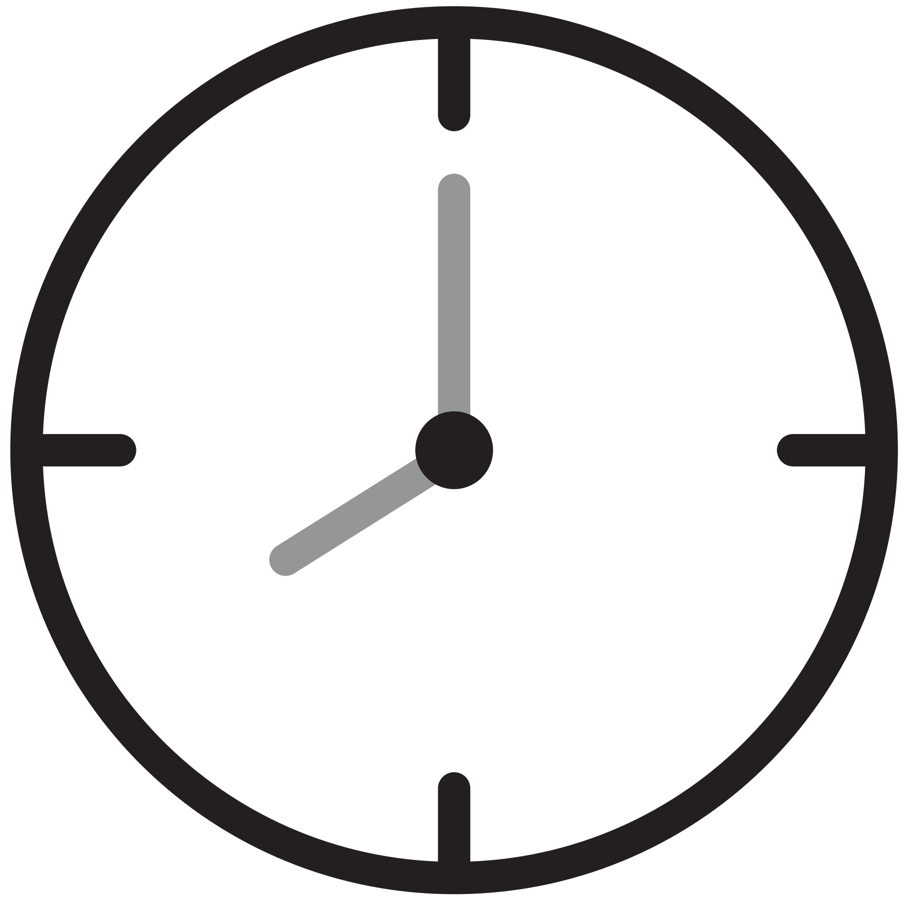


How have rates of evolution varied (or not) across the tree?

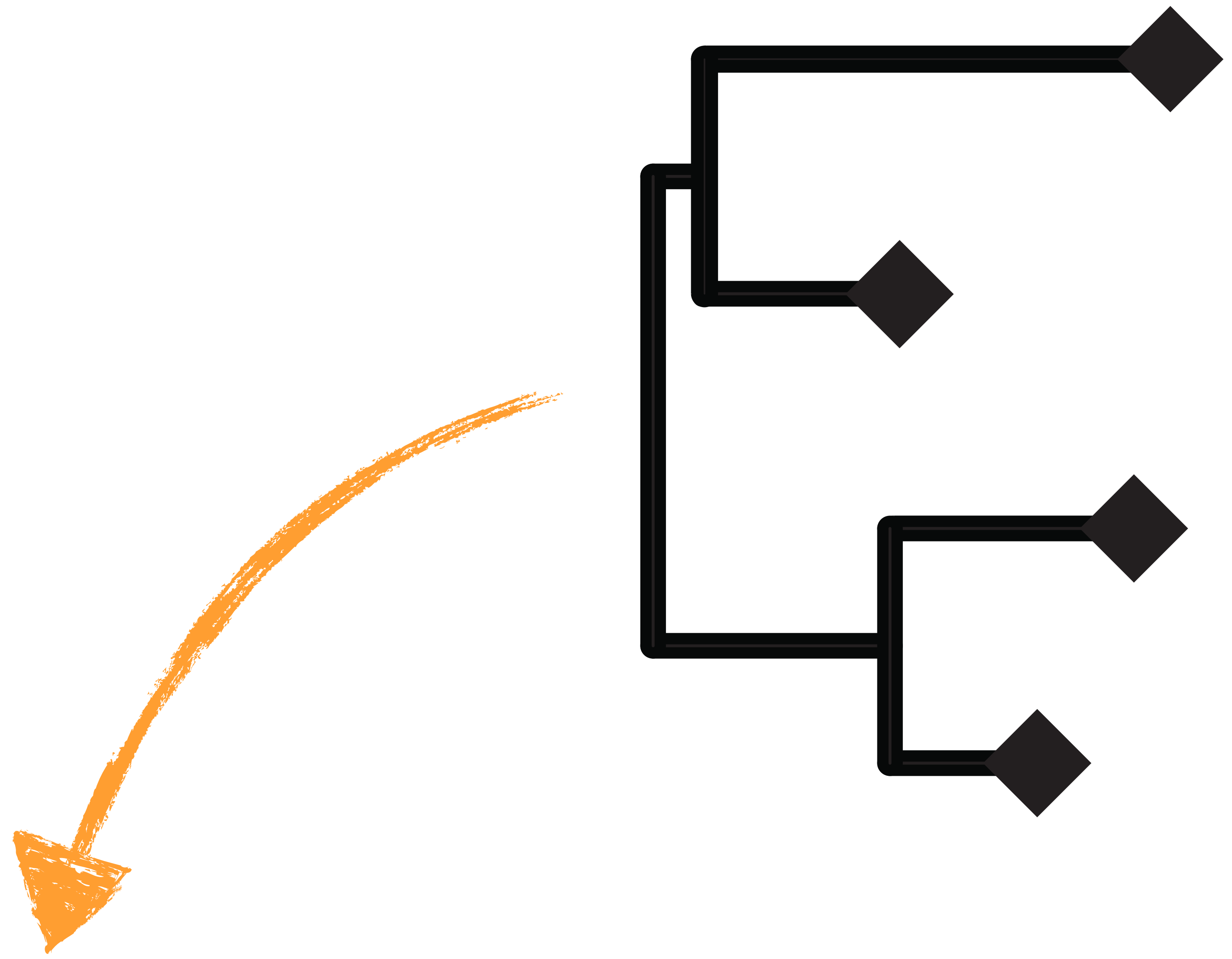
substitution model



clock model

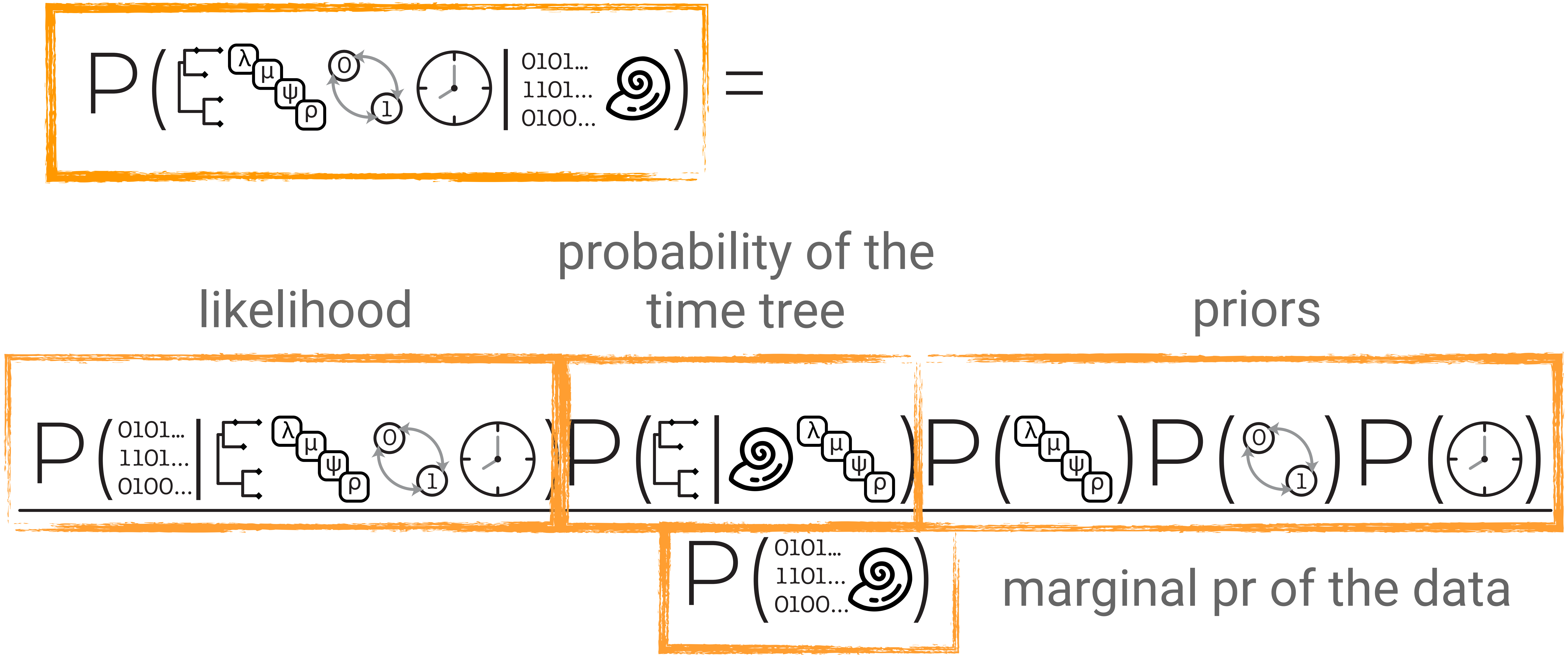


tree model



How have species originated, gone extinct and been sampled through time?

Bayesian divergence time estimation



posterior

likelihood

probability of the time tree

priors

marginal pr of the data

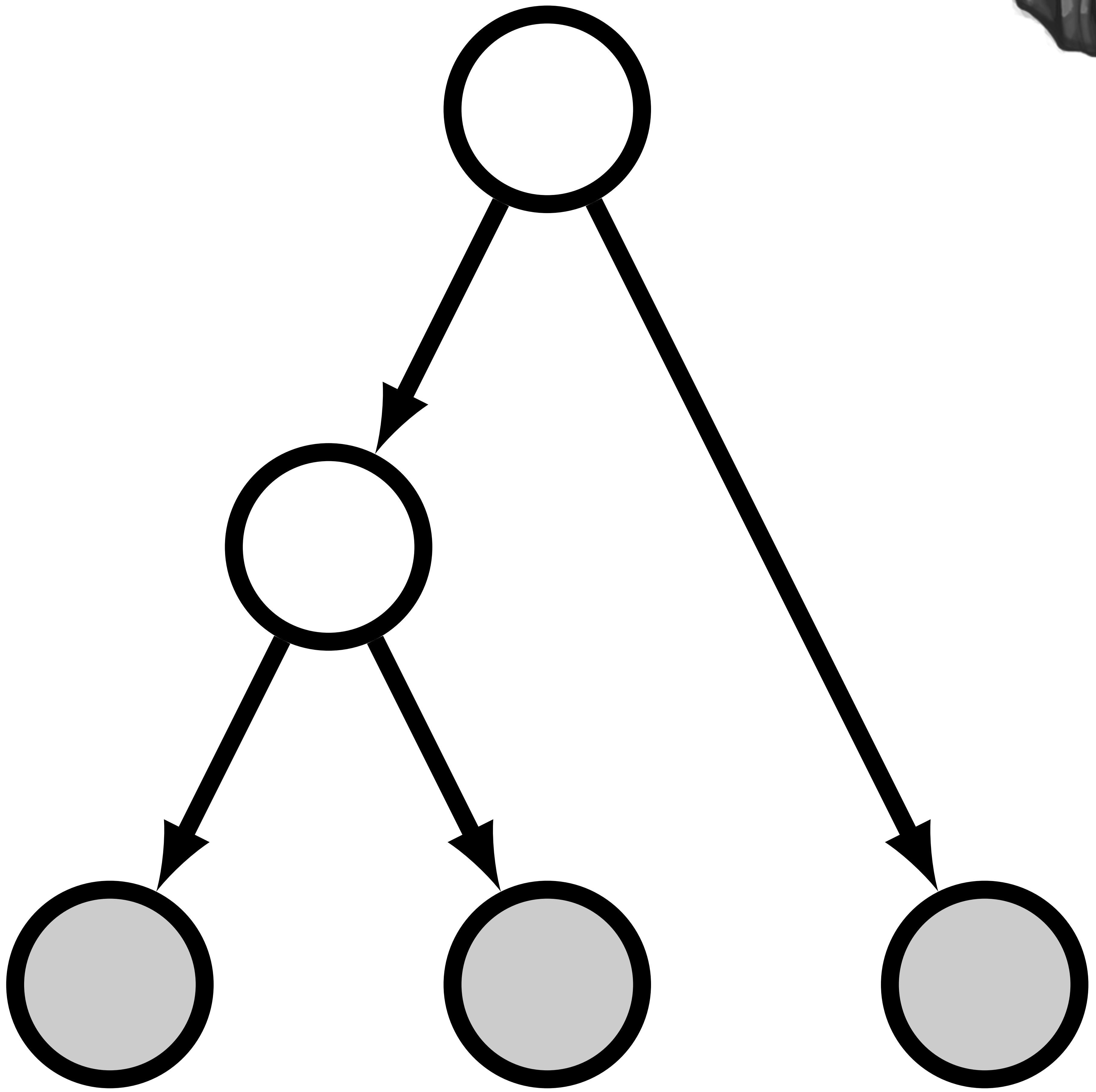
RevBayes

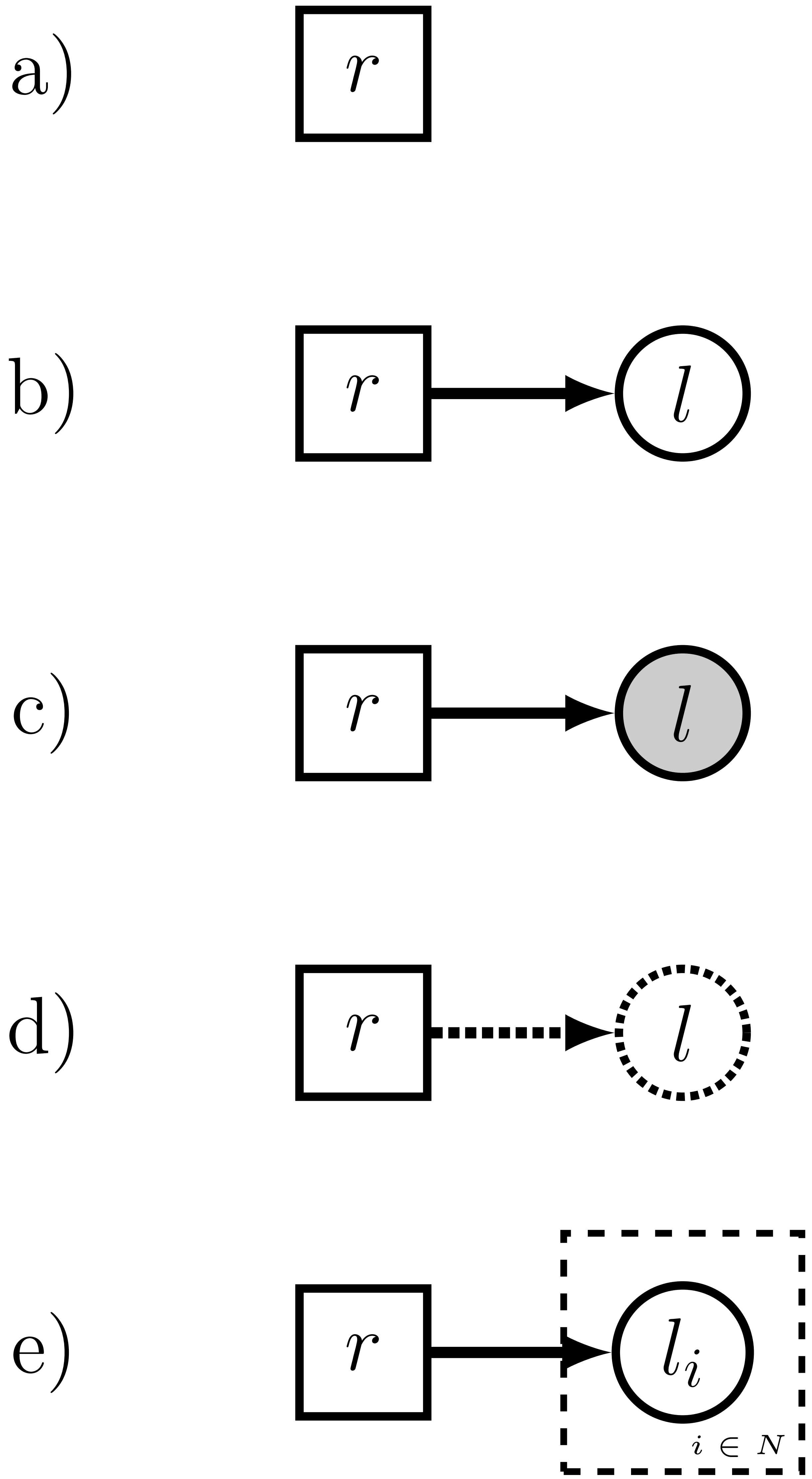
Designed with extendability and flexibility in mind

Rev language, similar to R, and uses a graphical modelling framework

Developed and supported by a large international team of developers

revbayes.github.io





```
# constant node
r <- 10


# stochastic node
l ∼ dnExp(r)


# stochastic node (observed)
l.clamp(0.1)


# deterministic node
l := exp(r)


# stochastic nodes (iid)
for (i in 1:N) {
    l[i] ∼ dnExp(r)
}
```

Intro to BEAST2

BEAST2

Beast2

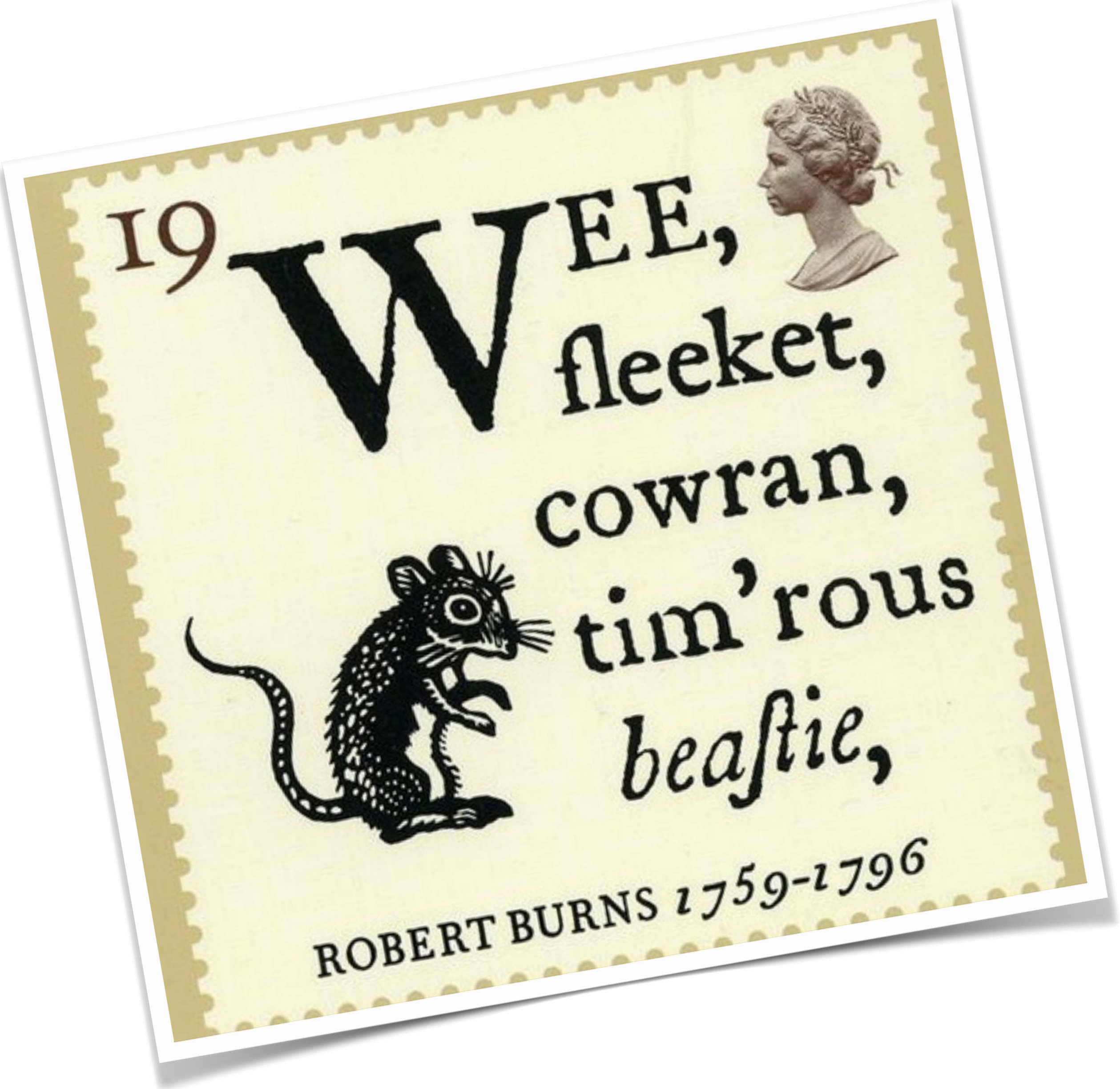
Bayesian evolutionary analysis by sampling trees

Also designed with extendability and flexibility in mind

Also developed and supported by a large international team of developers

Has a suite of apps that can used to generate input files and analysis the output

www.beast2.org



Scots poem - also the BEAST2 logo!

BEAST2 toolkit and work flow

Step 1. generate the xml file in **BEAUti**

Step 2. run your analysis in **BEAST**

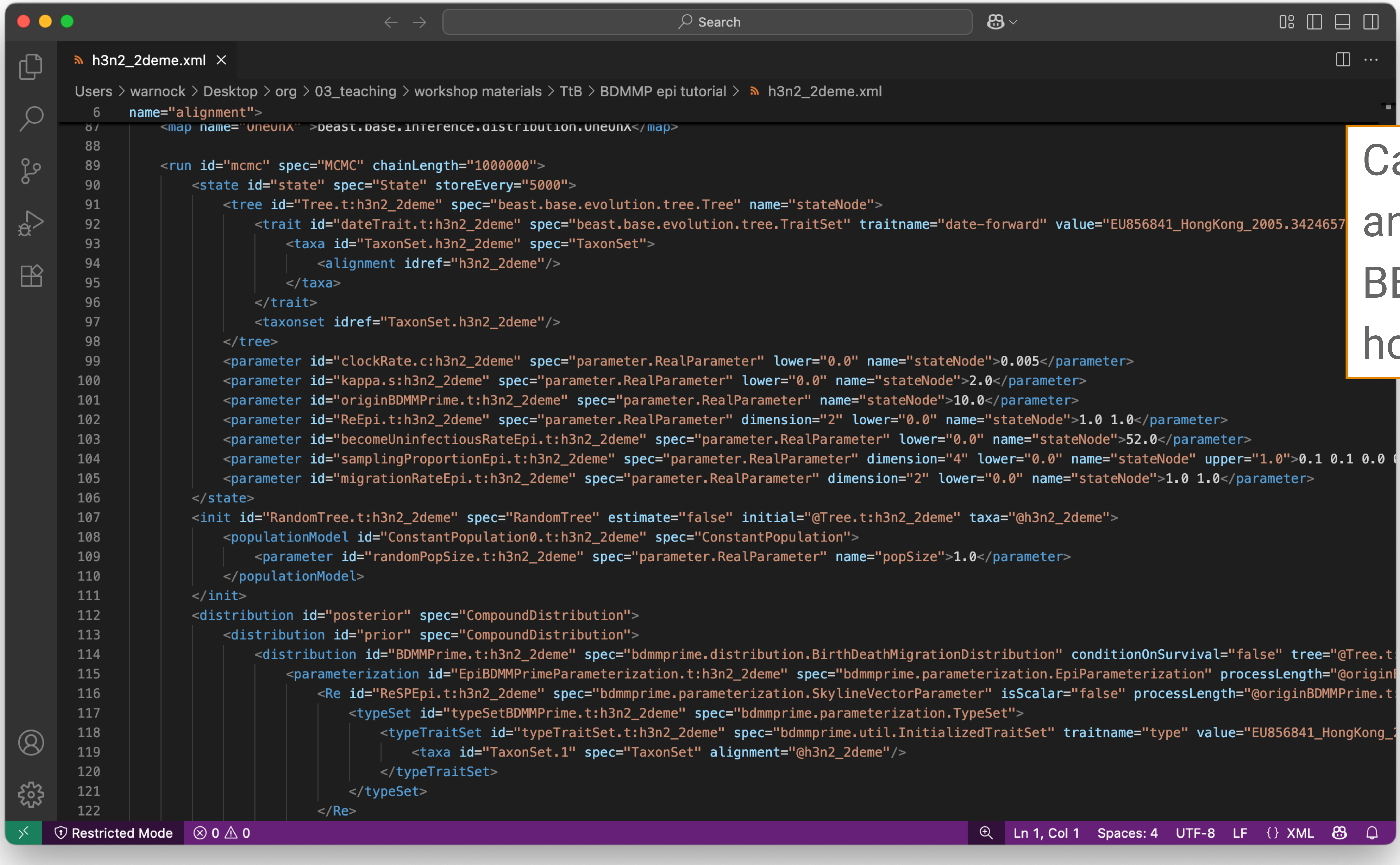
Step 3a. Examine you log files using **Tracer**

Step 3b. Generate a summary tree using **TreeAnnotator**

Step 4. Examine your summary tree in **FigTree**

Step… any other downstream analysis

BEAST2 input: the XML file



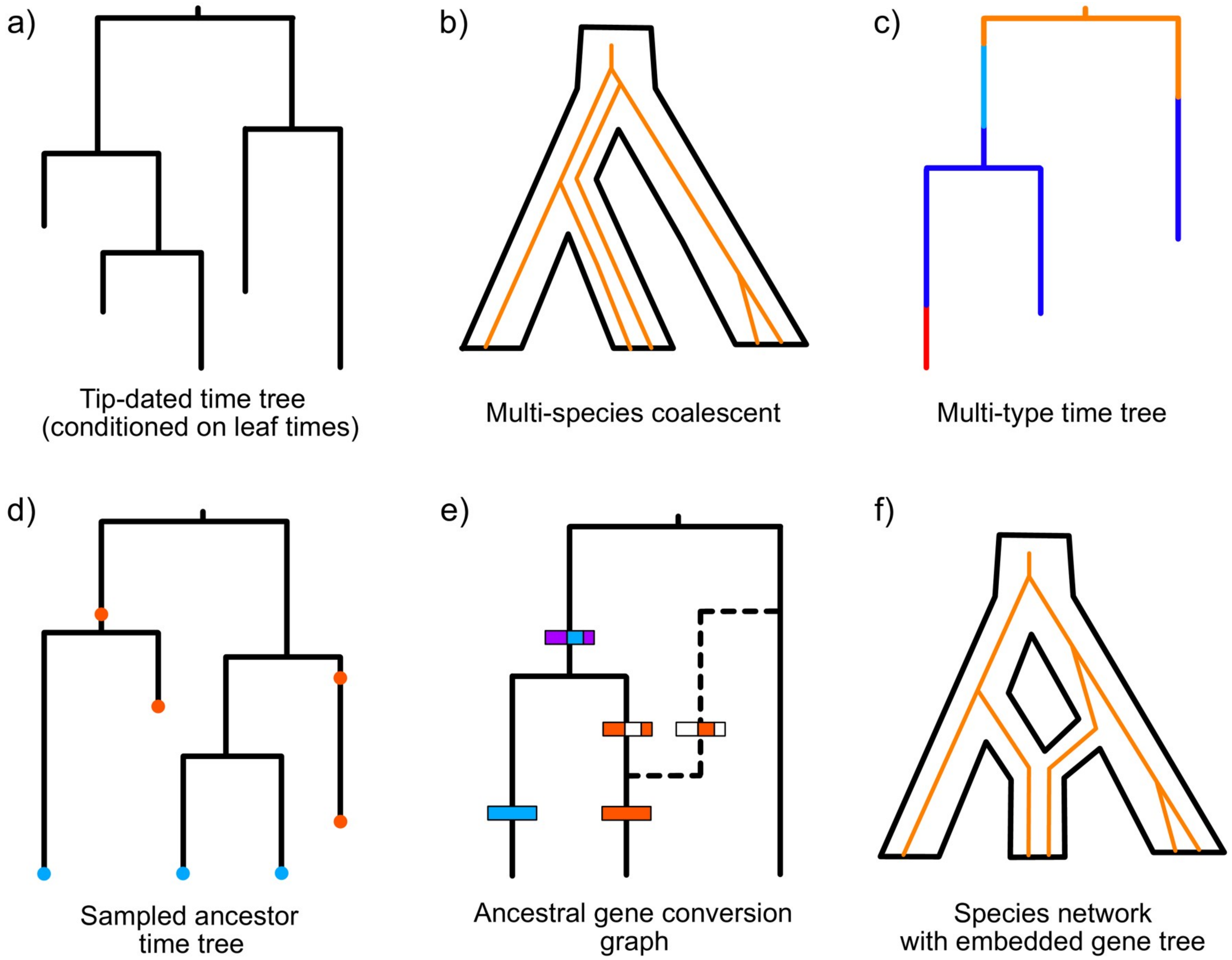
```xml
name="alignment">
    <map name="OneOnX" >beast.base.inference.distribution.OneOnX</map>

    <run id="mcmc" spec="MCMC" chainLength="1000000">
        <state id="state" spec="State" storeEvery="5000">
            <tree id="Tree.t:h3n2_2deme" spec="beast.base.evolution.tree.Tree" name="stateNode">
                <trait id="dateTrait.t:h3n2_2deme" spec="beast.base.evolution.tree.TraitSet" traitname="date-forward" value="EU856841_HongKong_2005.3424657
                    <taxa id="TaxonSet.h3n2_2deme" spec="TaxonSet">
                        <alignment idref="h3n2_2deme"/>
                    </taxa>
                </trait>
                <taxonset idref="TaxonSet.h3n2_2deme"/>
            </tree>
            <parameter id="clockRate.c:h3n2_2deme" spec="parameter.RealParameter" lower="0.0" name="stateNode">0.005</parameter>
            <parameter id="kappa.s:h3n2_2deme" spec="parameter.RealParameter" lower="0.0" name="stateNode">2.0</parameter>
            <parameter id="originBDMMPrime.t:h3n2_2deme" spec="parameter.RealParameter" name="stateNode">10.0</parameter>
            <parameter id="ReEpi.t:h3n2_2deme" spec="parameter.RealParameter" dimension="2" lower="0.0" name="stateNode">1.0 1.0</parameter>
            <parameter id="becomeUninfectiousRateEpi.t:h3n2_2deme" spec="parameter.RealParameter" lower="0.0" name="stateNode">52.0</parameter>
            <parameter id="samplingProportionEpi.t:h3n2_2deme" spec="parameter.RealParameter" dimension="4" lower="0.0" name="stateNode" upper="1.0">0.1 0.1 0.0
            <parameter id="migrationRateEpi.t:h3n2_2deme" spec="parameter.RealParameter" dimension="2" lower="0.0" name="stateNode">1.0 1.0</parameter>
        </state>
        <init id="RandomTree.t:h3n2_2deme" spec="RandomTree" estimate="false" initial="@Tree.t:h3n2_2deme" taxa="@h3n2_2deme">
            <populationModel id="ConstantPopulation0.t:h3n2_2deme" spec="ConstantPopulation">
                <parameter id="randomPopSize.t:h3n2_2deme" spec="parameter.RealParameter" name="popSize">1.0</parameter>
            </populationModel>
        </init>
        <distribution id="posterior" spec="CompoundDistribution">
            <distribution id="prior" spec="CompoundDistribution">
                <distribution id="BDMMPrime.t:h3n2_2deme" spec="bdmmprime.distribution.BirthDeathMigrationDistribution" conditionOnSurvival="false" tree="@Tree.t
                    <parameterization id="EpiBDMMPrimeParameterization.t:h3n2_2deme" spec="bdmmprime.parameterization.EpiParameterization" processLength="@origin
                        <Re id="ReSPEpi.t:h3n2_2deme" spec="bdmmprime.parameterization.SkylineVectorParameter" isScalar="false" processLength="@originBDMMPrime.t
                            <typeSet id="typeSetBDMMPrime.t:h3n2_2deme" spec="bdmmprime.parameterization.TypeSet">
                                <typeTraitSet id="typeTraitSet.t:h3n2_2deme" spec="bdmmprime.util.InitializedTraitSet" traitname="type" value="EU856841_HongKong_
                                    <taxa id="TaxonSet.1" spec="TaxonSet" alignment="@h3n2_2deme"/>
                                </typeTraitSet>
                            </typeSet>
                        </Re>
```

Caveat: if you want to use an option not available in BEAUti you have to learn how to edit the XML

A wide range of models and tree structures

Note all tree models in BEAST2 incorporate a temporal component



Tip-dated time tree
(conditioned on leaf times)

Multi-species coalescent

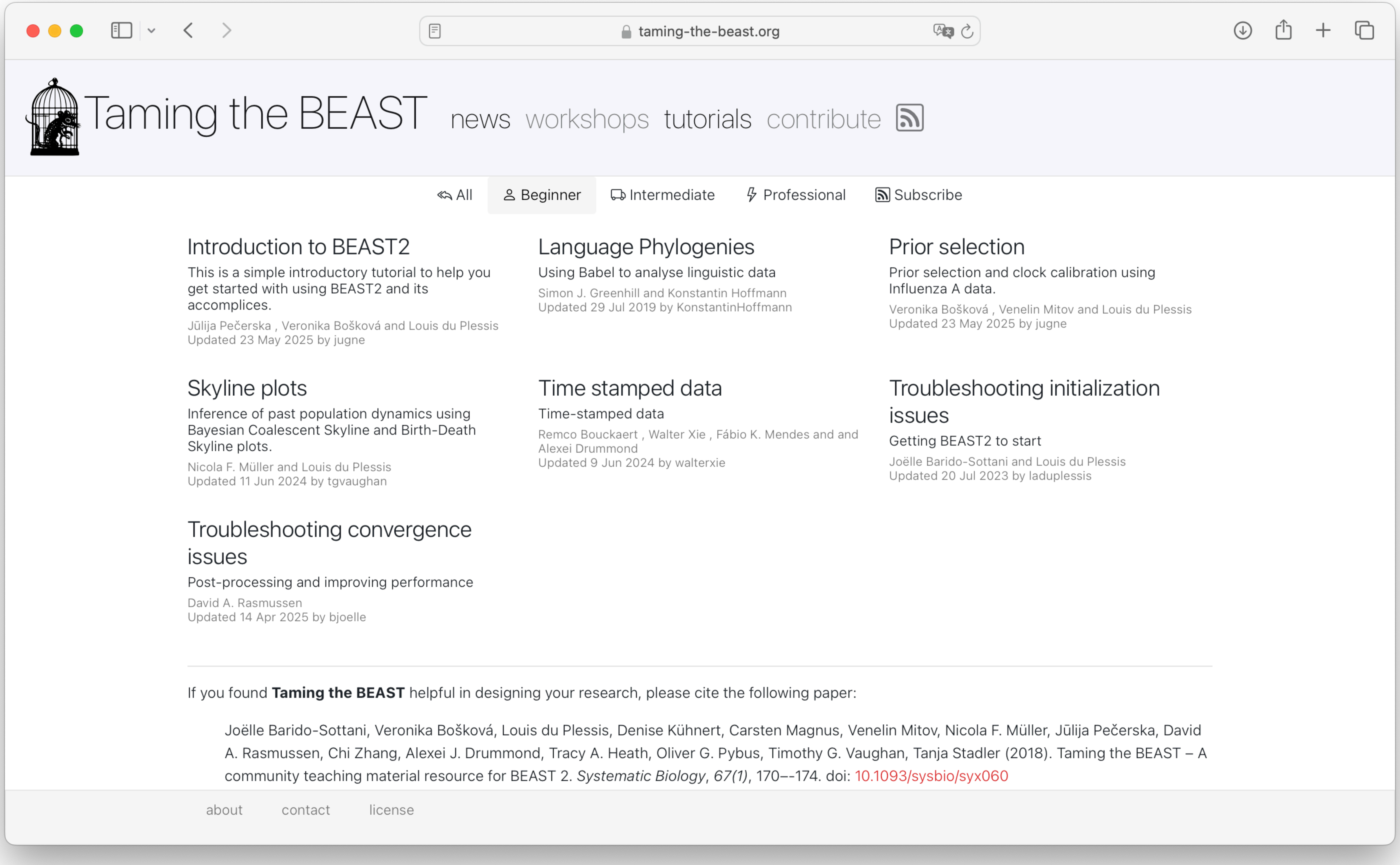
Multi-type time tree

Sampled ancestor
time tree

Ancestral gene conversion
graph

Species network
with embedded gene tree

Bouckaert et al. (2019) PLOS Comp Bio



Introduction to BEAST2

This is a simple introductory tutorial to help you get started with using BEAST2 and its accomplices.

Jūlija Pečerska , Veronika Bošková and Louis du Plessis
Updated 23 May 2025 by jugne

Language Phylogenies

Using Babel to analyse linguistic data

Simon J. Greenhill and Konstantin Hoffmann
Updated 29 Jul 2019 by KonstantinHoffmann

Prior selection

Prior selection and clock calibration using Influenza A data.

Veronika Bošková , Venelin Mitov and Louis du Plessis
Updated 23 May 2025 by jugne

Skyline plots

Inference of past population dynamics using Bayesian Coalescent Skyline and Birth-Death Skyline plots.

Nicola F. Müller and Louis du Plessis
Updated 11 Jun 2024 by tgvaughan

Time stamped data

Time-stamped data

Remco Bouckaert , Walter Xie , Fábio K. Mendes and and Alexei Drummond
Updated 9 Jun 2024 by walterxie

Troubleshooting initialization issues

Getting BEAST2 to start

Joëlle Barido-Sottani and Louis du Plessis
Updated 20 Jul 2023 by laduplessis

Troubleshooting convergence issues

Post-processing and improving performance

David A. Rasmussen
Updated 14 Apr 2025 by bjoelle

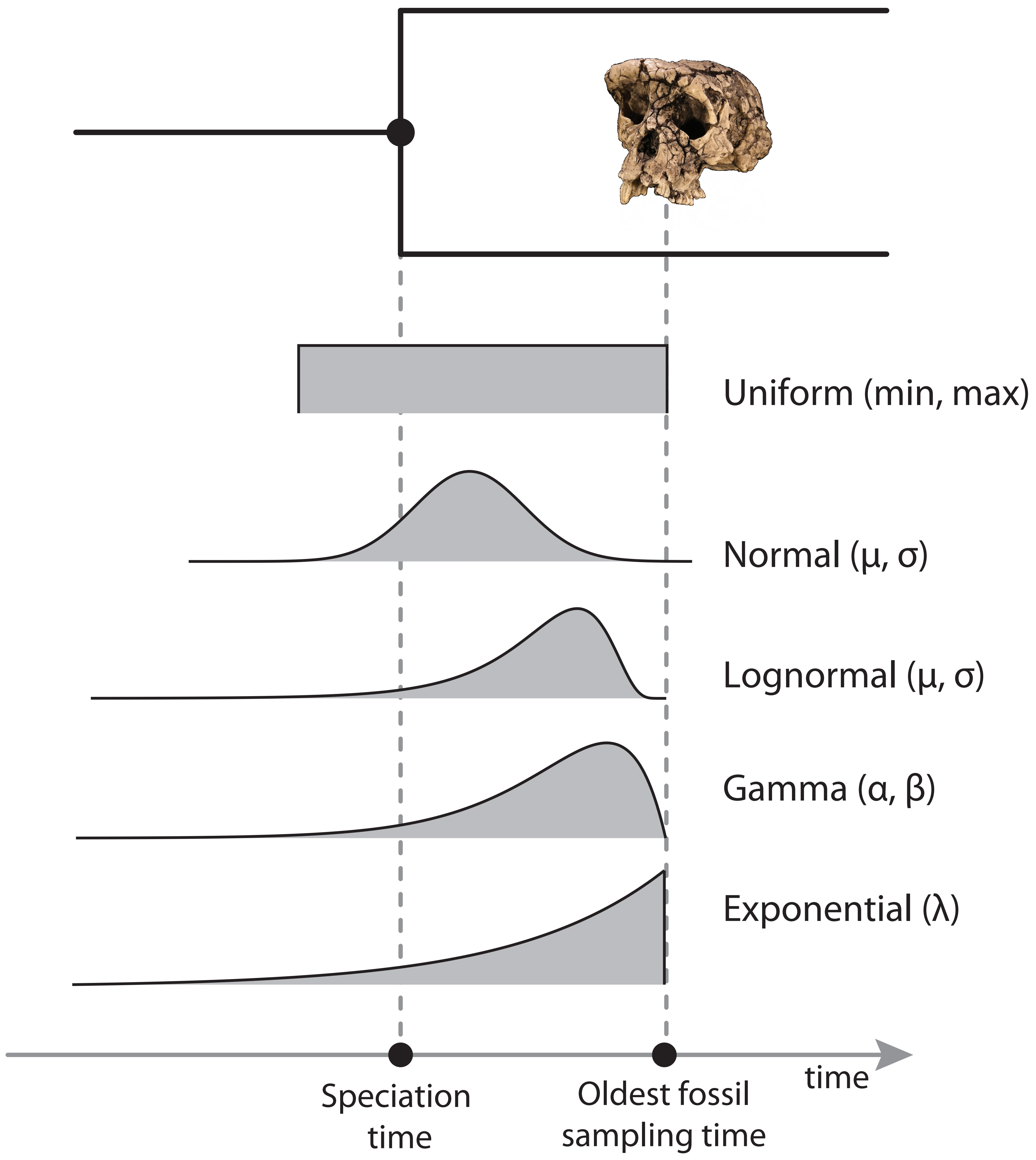
If you found **Taming the BEAST** helpful in designing your research, please cite the following paper:

> Joëlle Barido-Sottani, Veronika Bošková, Louis du Plessis, Denise Kühnert, Carsten Magnus, Venelin Mitov, Nicola F. Müller, Jūlija Pečerska, David A. Rasmussen, Chi Zhang, Alexei J. Drummond, Tracy A. Heath, Oliver G. Pybus, Timothy G. Vaughan, Tanja Stadler (2018). Taming the BEAST – A community teaching material resource for BEAST 2. *Systematic Biology*, *67(1)*, 170–-174. doi: 10.1093/sysbio/syx060

https://taming-the-beast.org/

Recap: node dating

Node dating

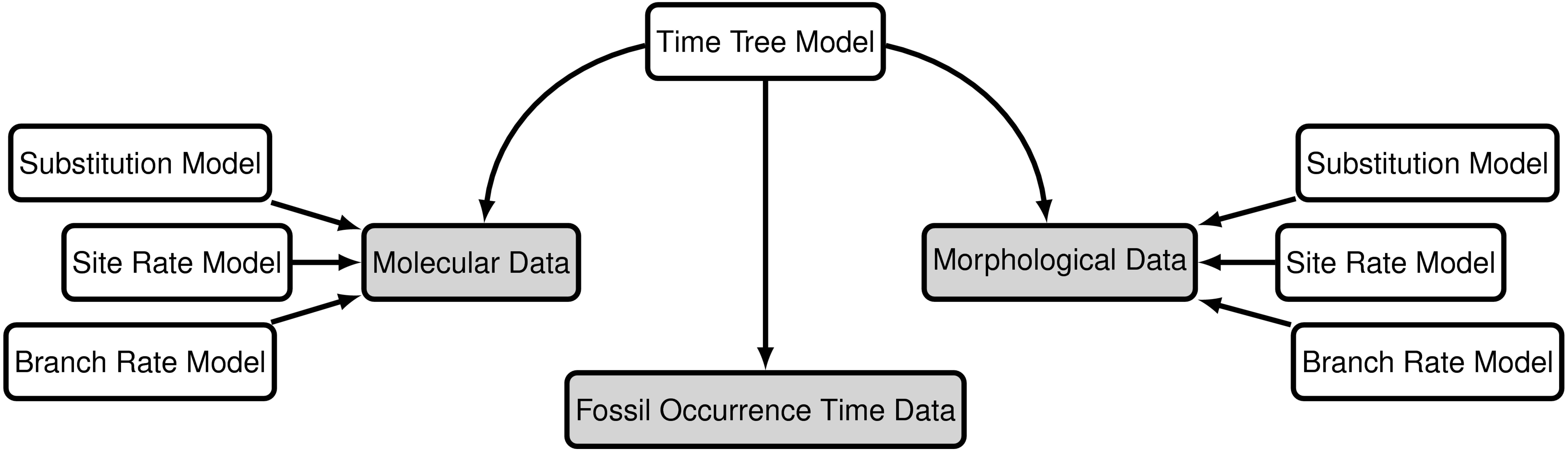


We can use a calibration density to constrain internal node ages

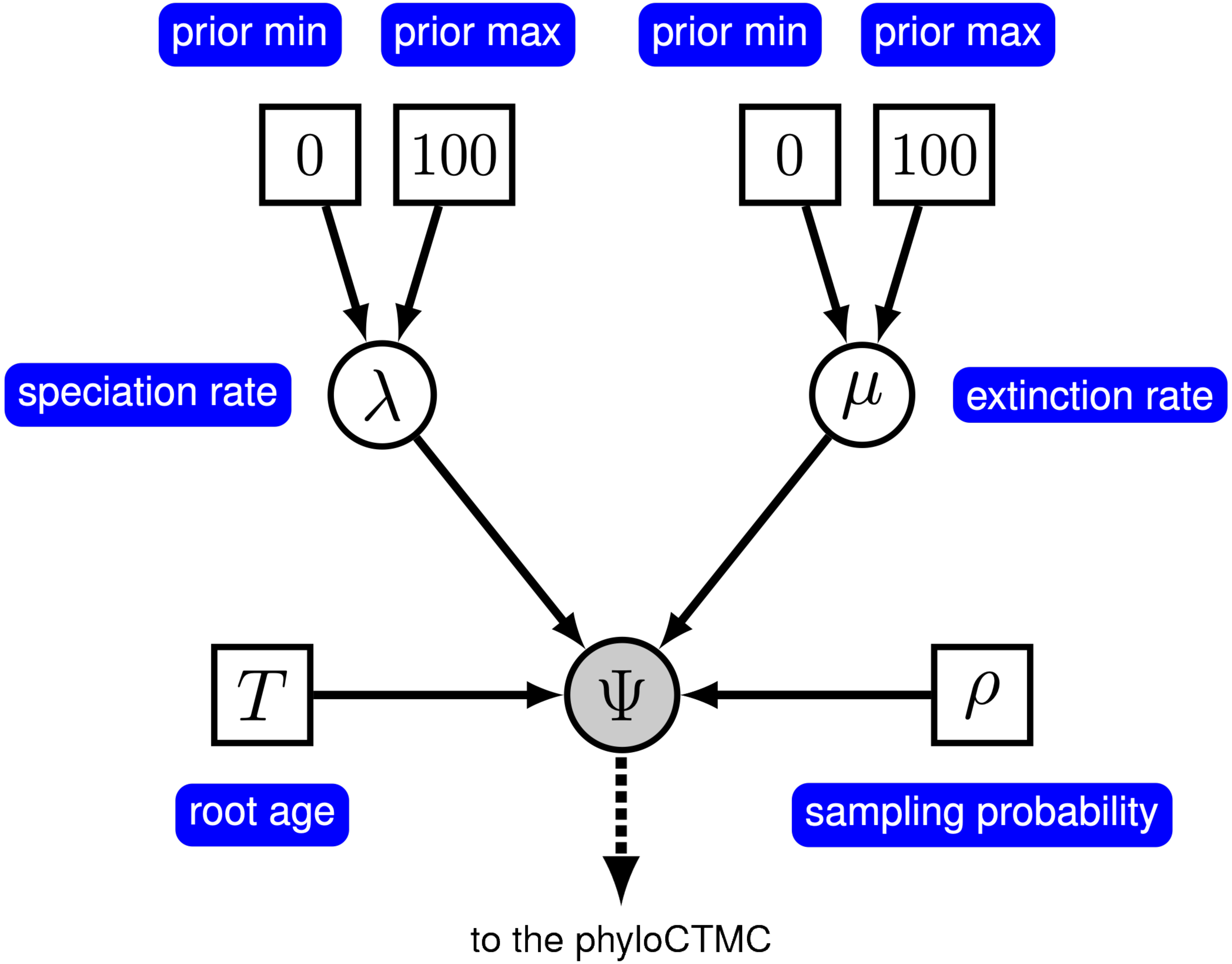
We typically use a birth-death process model to describe the tree generating process

Adapted from *Heath* (2012) *Sys Bio*

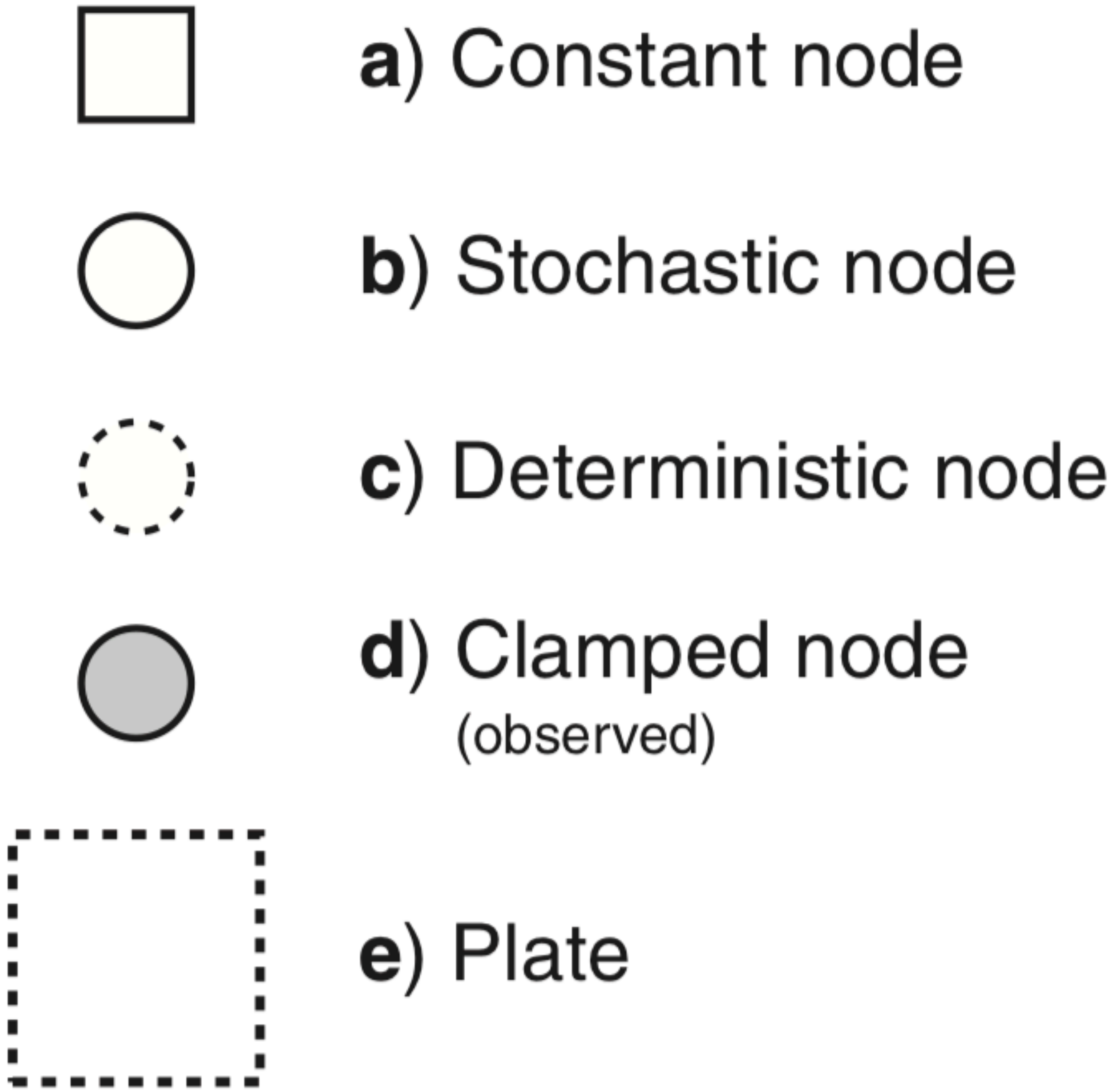
Graphical representation of the tripartite model



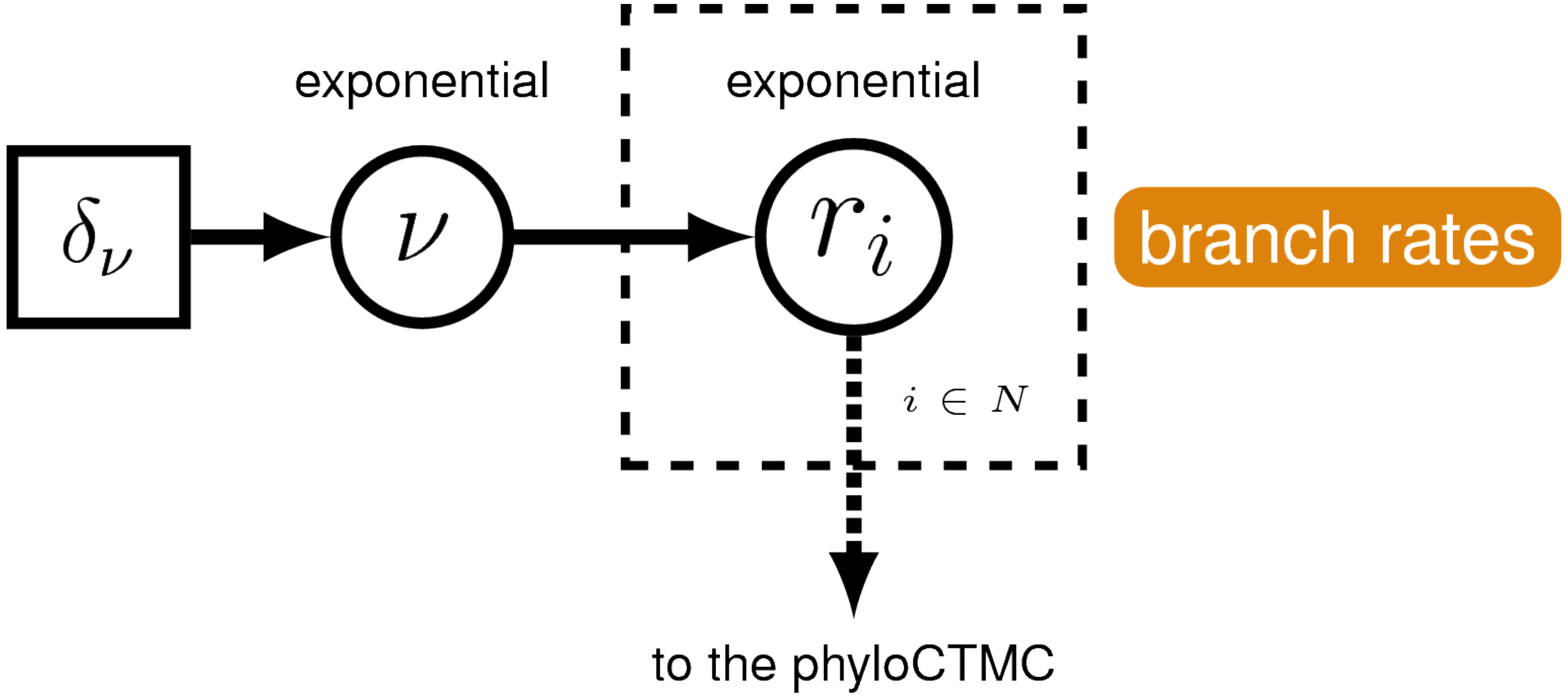
Graphical representation of the birth-death process



Graphical models: relaxed clock model



- **a**) Constant node
- **b**) Stochastic node
- **c**) Deterministic node
- **d**) Clamped node
  (observed)
- **e**) Plate



Exercise

Scenario	**Software**
Node dating with large datasets	MCMCTree
If want (or have to) fix the tree topology	MCMCTree
If fossil sampling is sparse or complex	MCMCTree
If you have abundant fossil data, or are interested in the topological position of fossils	BEAST2, RevBayes
If you're interested in the phylodynamic parameters	BEAST2, RevBayes
If you want to use a specific model	BEAST2, RevBayes, MCMCTree

Recap

Bayes' theorem

$$\Pr(\textbf{model} \mid \textbf{data}) = \frac{\Pr(\textbf{data} \mid \textbf{model})\ \Pr(\textbf{model})}{\Pr(\textbf{data})}$$

Bayes' theorem

The probability of the data given the model assumptions and parameter values

Likelihood

$$\Pr(\text{model} \mid \text{data}) = \frac{\Pr(\text{data} \mid \text{model})\,\Pr(\text{model})}{\Pr(\text{data})}$$

Bayes’ theorem

Priors

This represents our prior knowledge of the model parameters

$$\Pr(\text{model} \mid \text{data}) = \frac{\Pr(\text{data} \mid \text{model})\,\Pr(\text{model})}{\Pr(\text{data})}$$

Bayes' theorem

$$\Pr(\textbf{model} \mid \textbf{data}) = \frac{\Pr(\textbf{data} \mid \textbf{model}) \Pr(\textbf{model})}{\Pr(\textbf{data})}$$

Marginal probability

The probability of the data, given all possible parameter values. Can be thought of as a normalising constant

Bayes' theorem

Reflects our combined knowledge based on the likelihood and the priors

posterior

$$\Pr(\textbf{model} \mid \textbf{data}) = \frac{\Pr(\textbf{data} \mid \textbf{model}) \Pr(\textbf{model})}{\Pr(\textbf{data})}$$

Bayesian tree inference