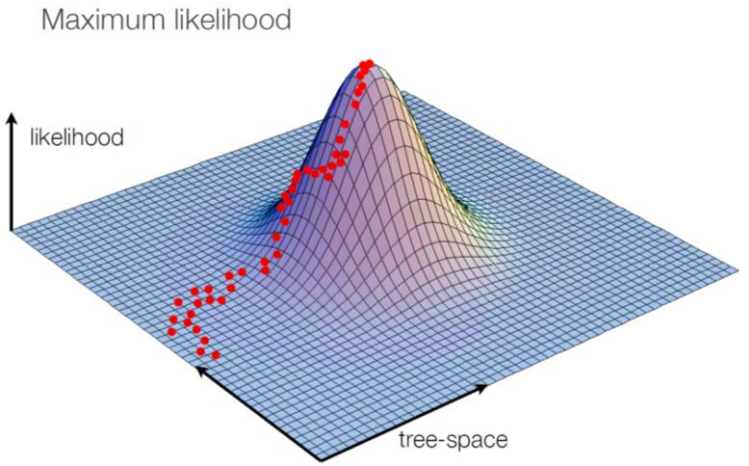
Day 3
Phylogenetic analysis using BEAST 2

Workshop

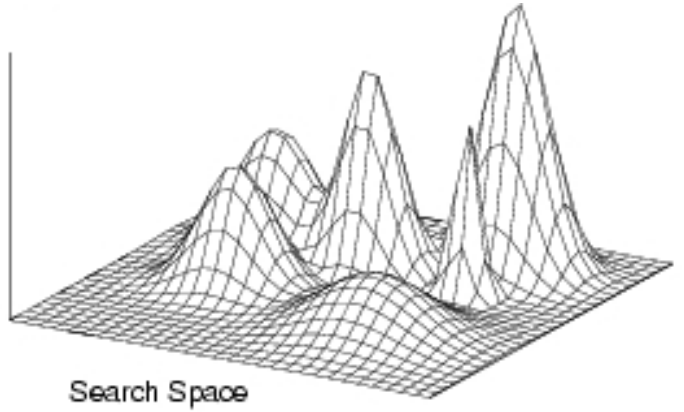
14.01.2020 – 17.01.2020

Search for the “true tree”

- Phylogenetic methods that use clustering algorithm (UPGMA,NJ) or Optimization methods (MP, ML) try to get the tree that best explains the data.

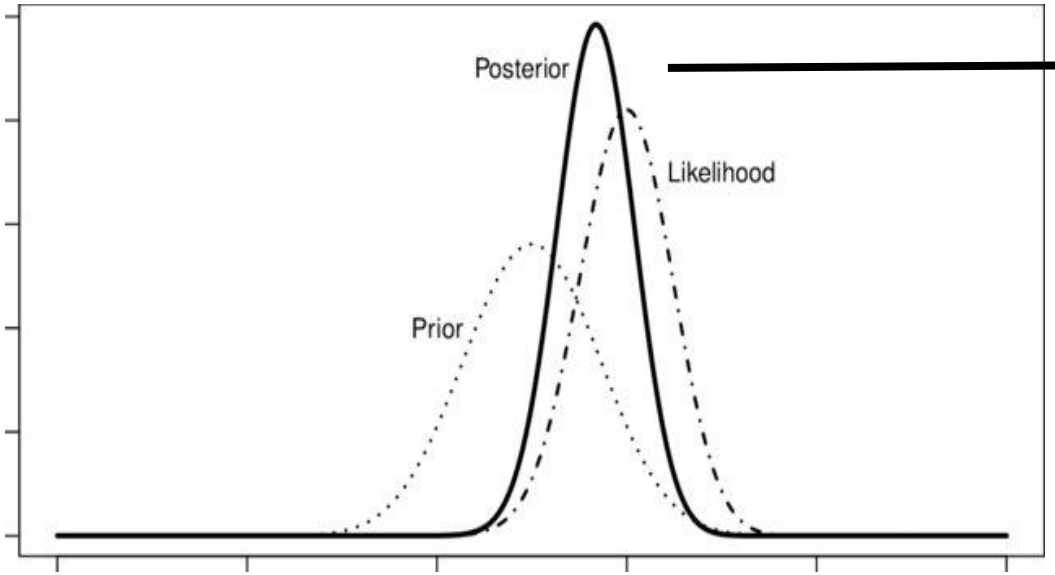


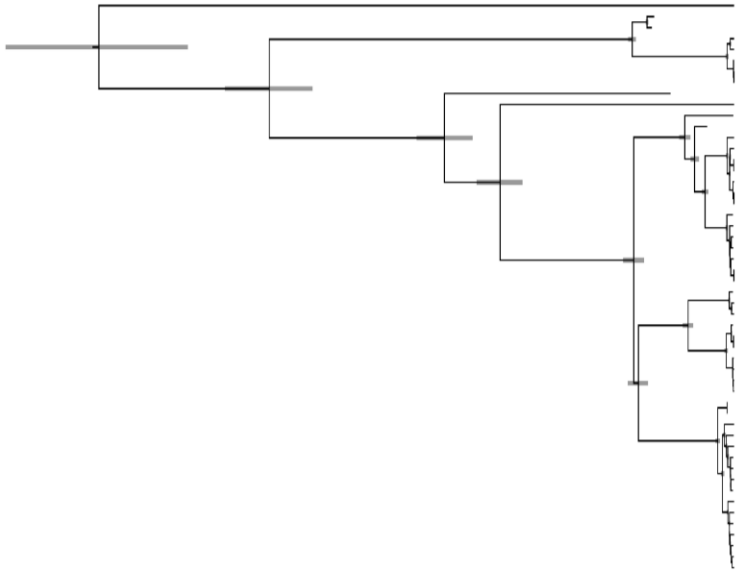
- Many trees can be as good or better:
  20 seq produce more than 8.2x10^20 rooted topologies.



Bayesian phylogenetics

- Allows to get a set of probable trees.
- Produces a posterior probability distribution.
- The posterior distribution can be translated to the probability of any branching event.





Bayesian inference

- Use to assess how our belief in a hypothesis (H) changes as a result of observing the data (D).

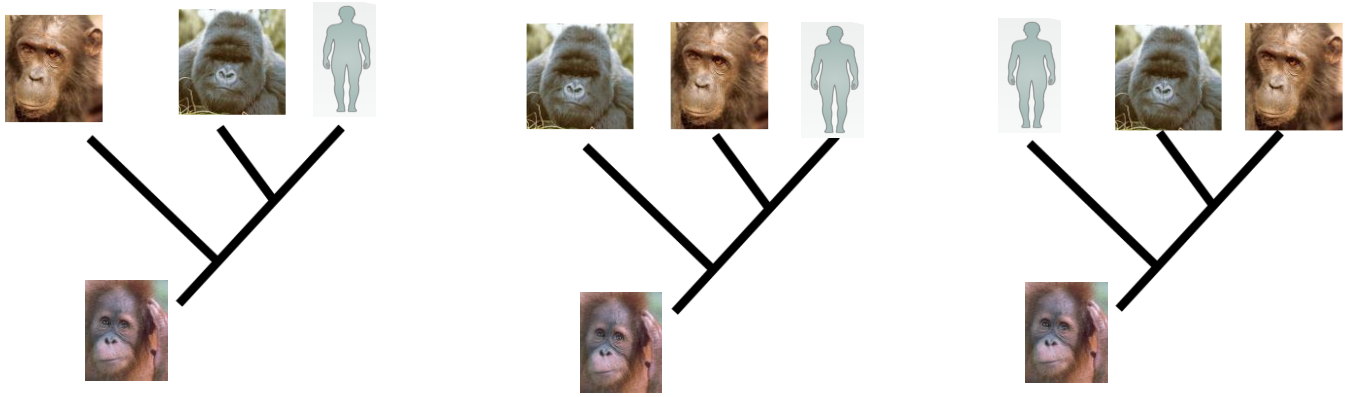
- **Bayes rule:** $\Pr(H|D) = \frac{\Pr(D|H)\Pr(H)}{\Pr(D)} \longrightarrow f(\tau|X) = \frac{f(X|\tau)f(\tau)}{f(X)}$

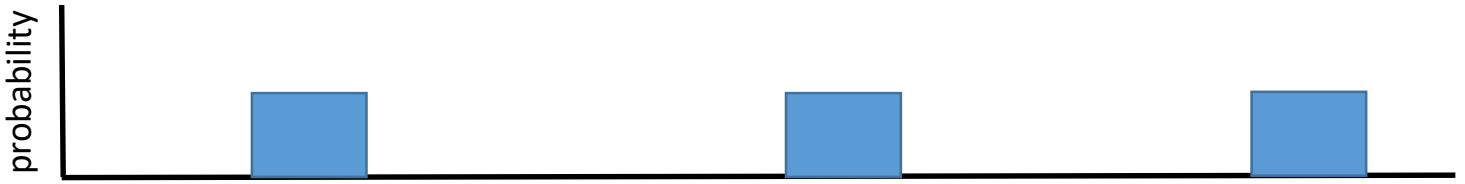
$f(X|\tau)$= probability of observing the data X if tree τ is true.

$f(\tau)$ = prior probability of tree τ

$$f(X) = \sum_{j=1}^{n} f(X|\tau_j)\, f(\tau_j)$$

Bayesian phylogenetic inference

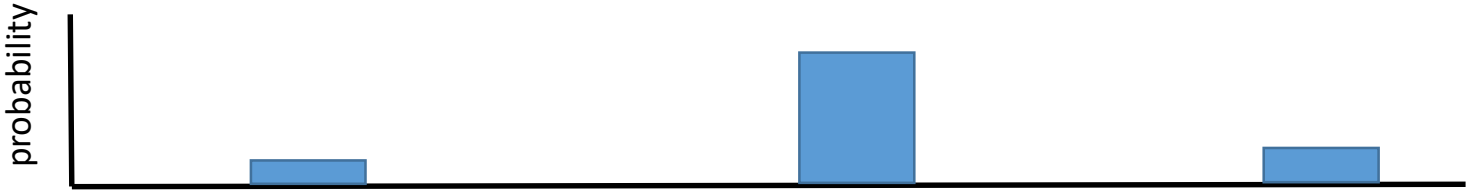




Prior distribution

Data

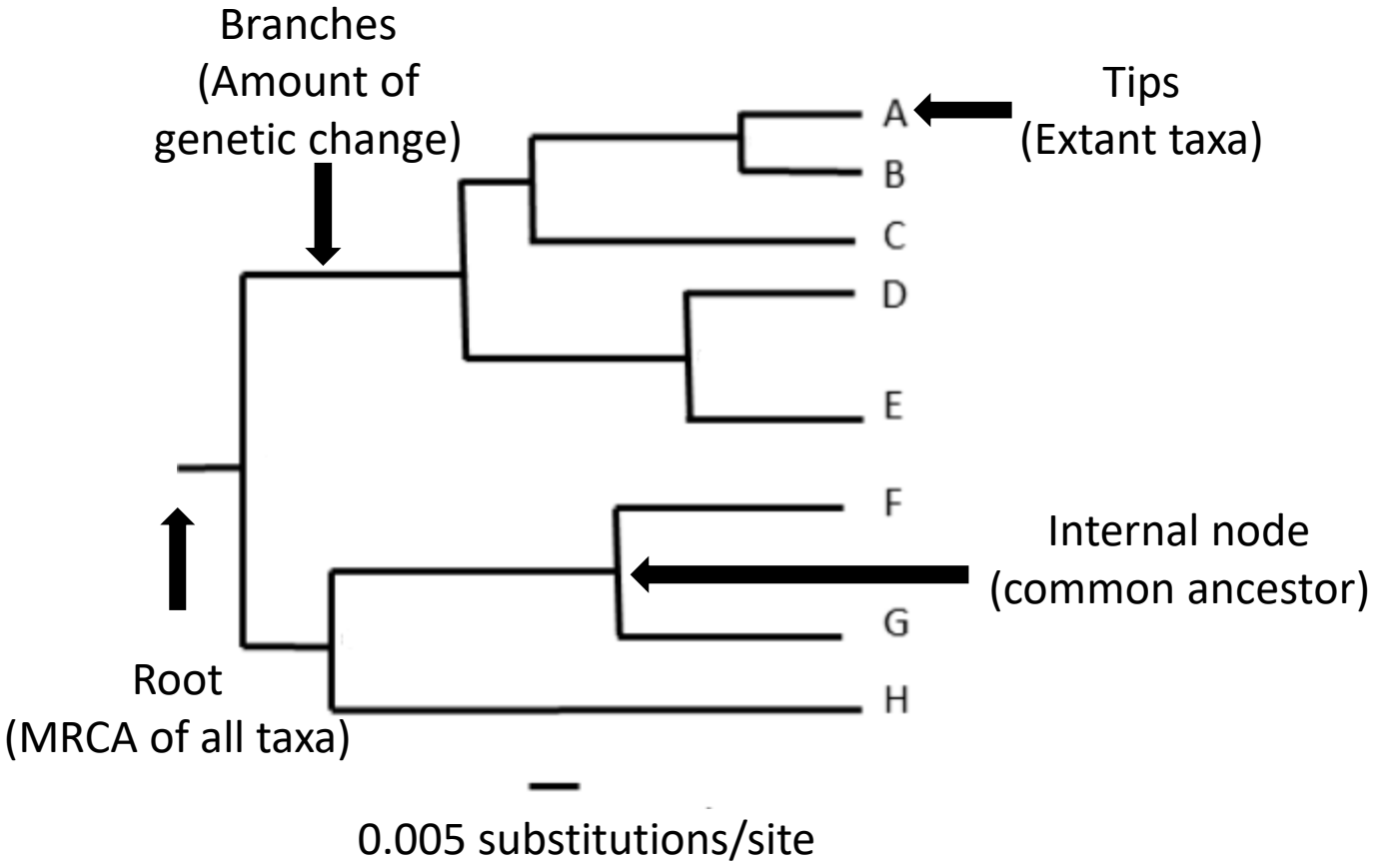
ACTGGGTCA
AGTGGCACC
AGCTGCAAC
AGCTGACCC



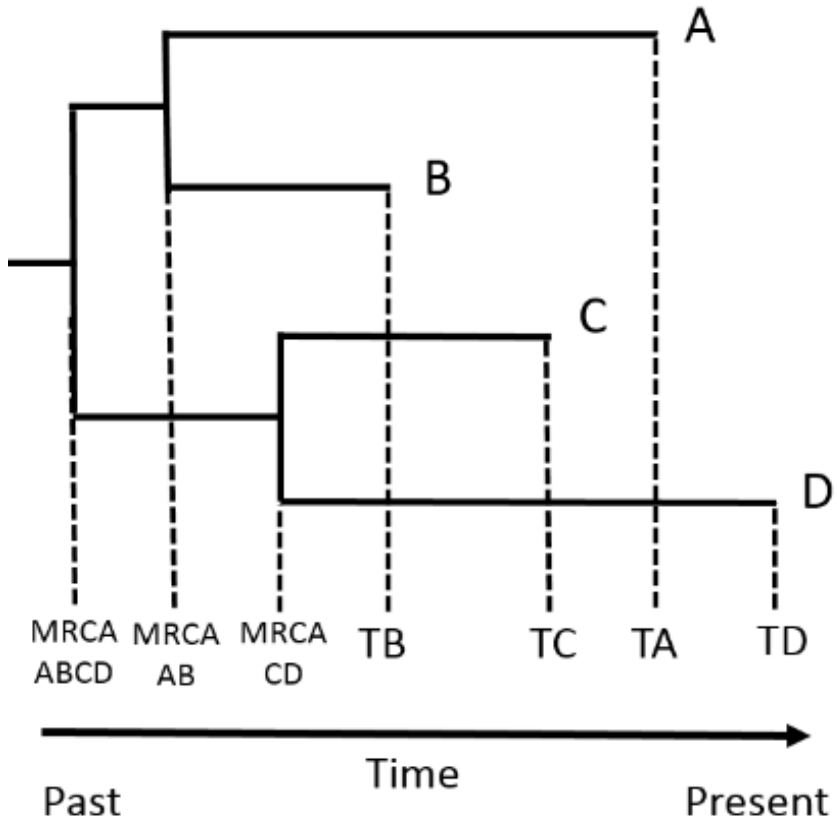
Posterior distribution

Time calibrated phylogenies

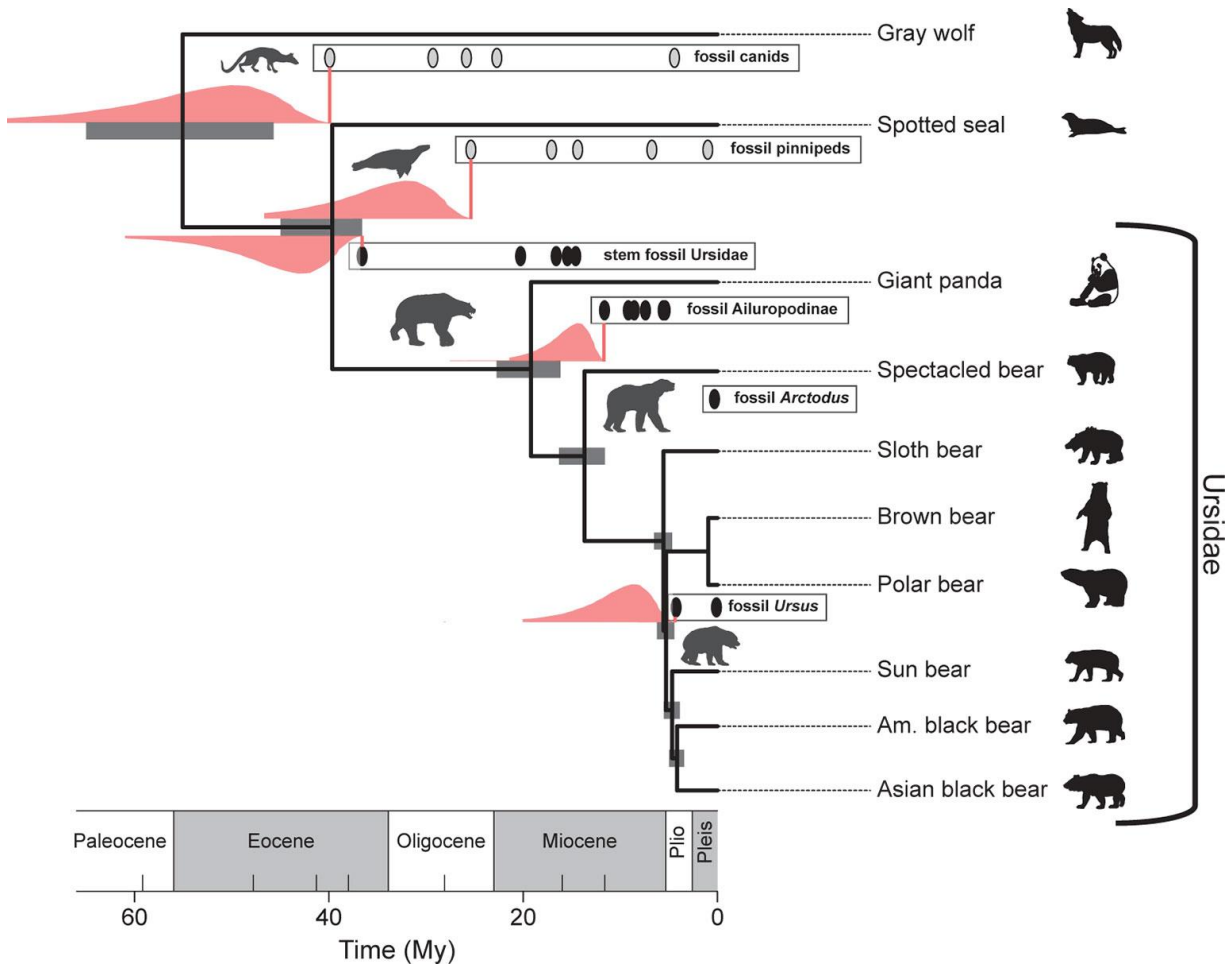
Rooted phylogenetic tree



Time-tree

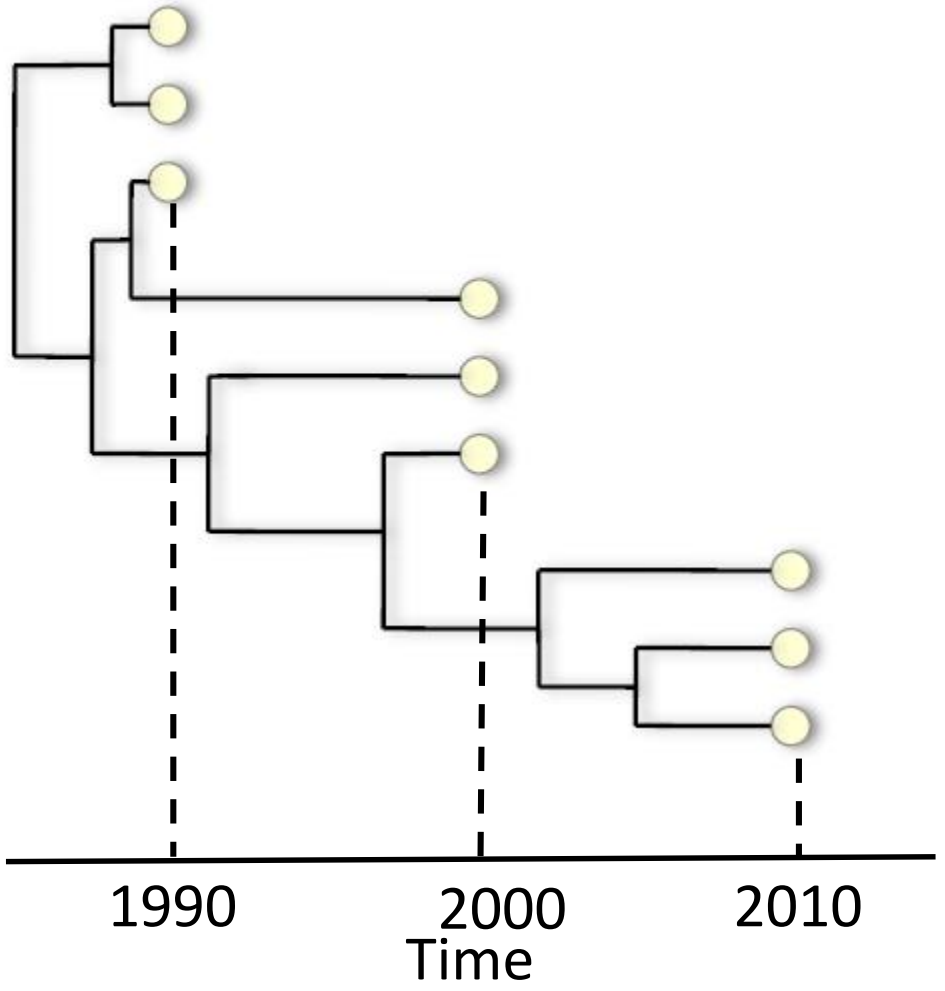


Node calibration



Marshall 2019

Tip dating



Elements of a Bayesian Phylogenetic Analysis

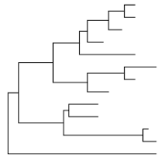
Evolutionary model

+

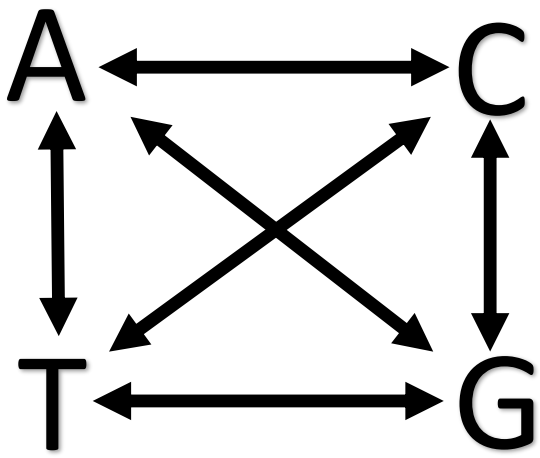
Rate heterogeneity among sites

Molecular clock

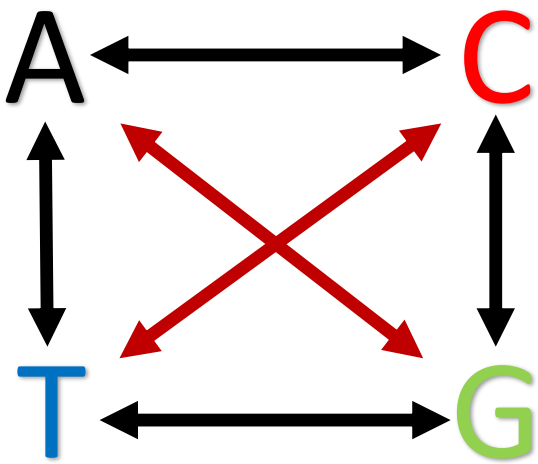
Tree Prior



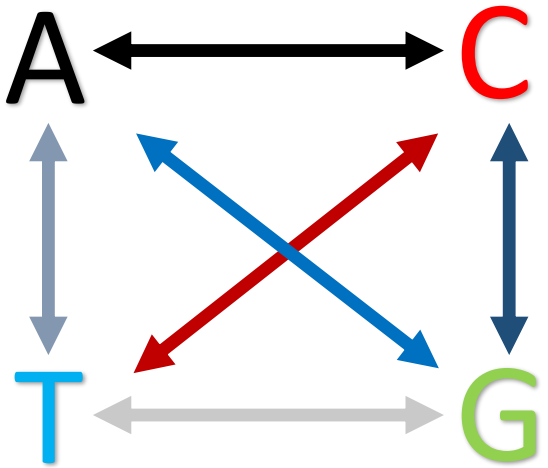
Evolutionary model



Jukes-Cantor 1969 (JC69)



Hasegawa-Kishino-Yano 1985 (HKY85)

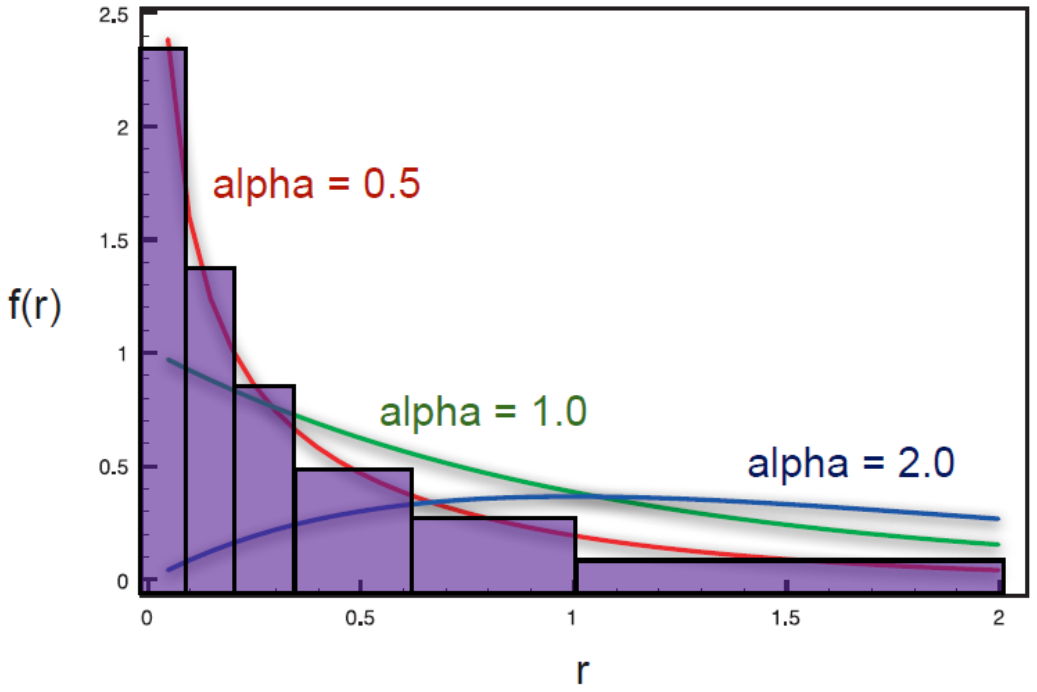


General Time Reversible(GTR)

Among-site rate heterogeneity

- Different genome regions evolve at different rates.
- Gamma model:

1. Allows different rates along the sequence
2. Modeled with 4-8 discrete rate categories
3. Shape parameter α

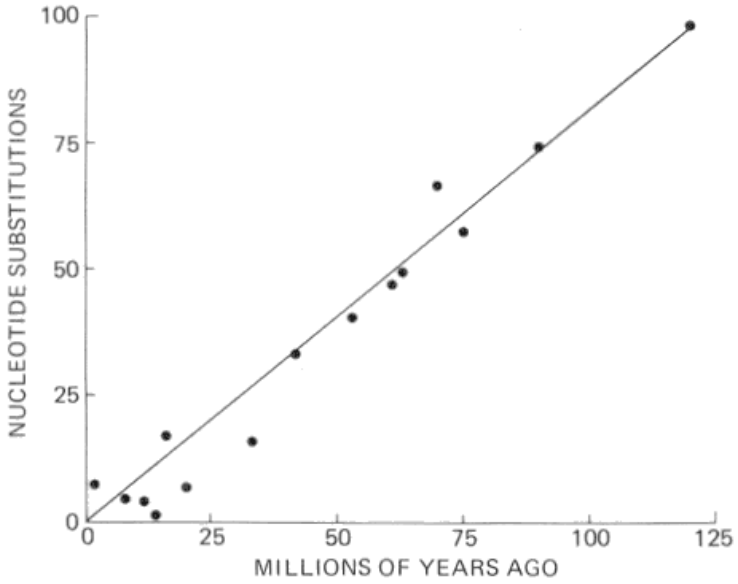


(Strict) Molecular clock

- Predicts a constant rate of molecular evolution among lineages.

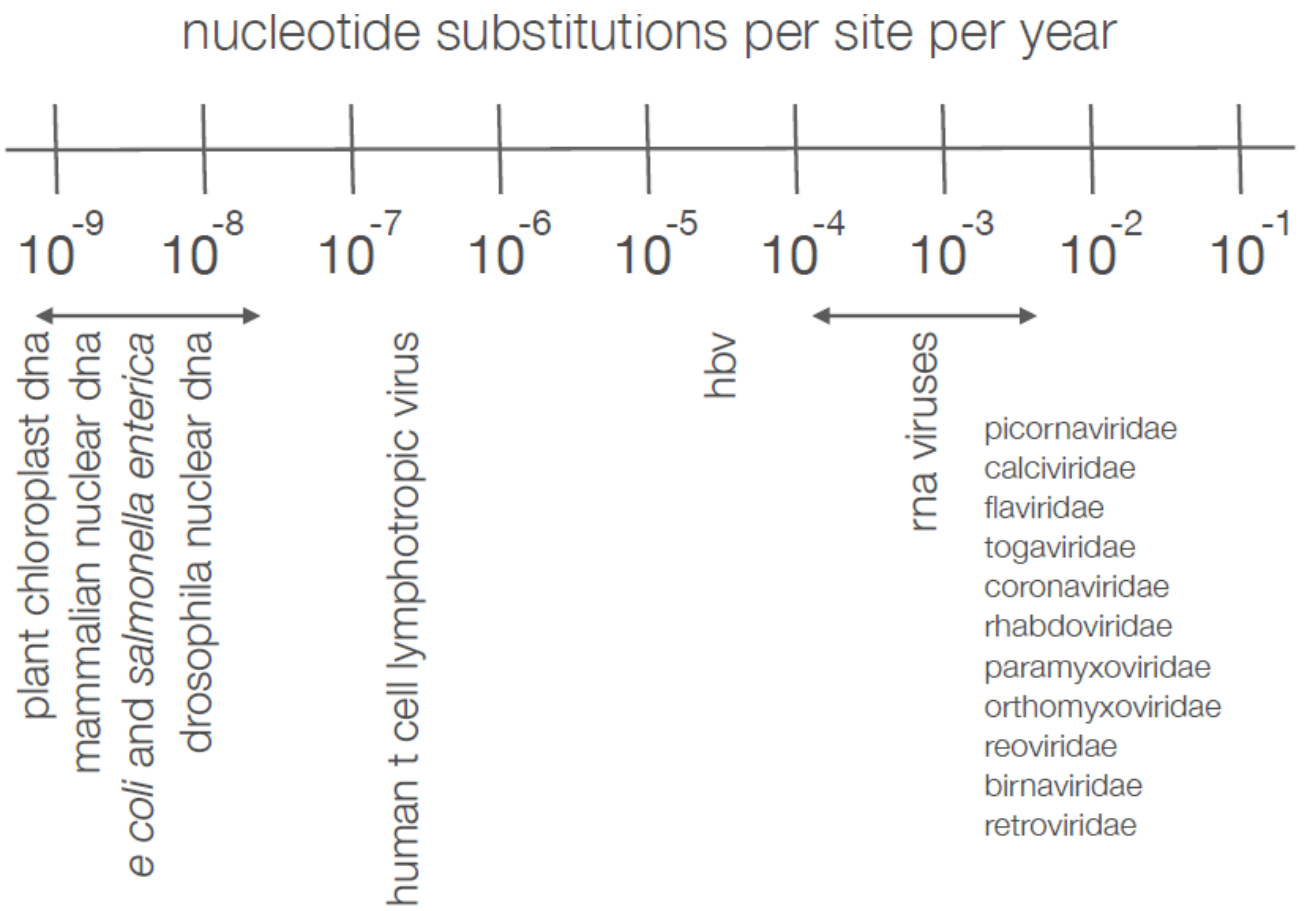
- Allows the estimation of evolutionary timescales.

- Molecular differences between pairs of species are proportional to the time of their separation.



Wilson 1976 Ho 2015

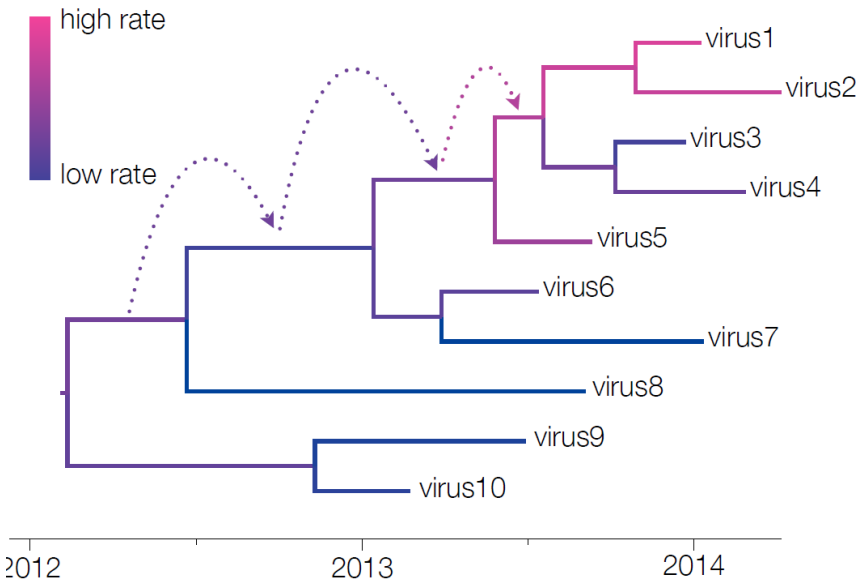
Different rates of evolution



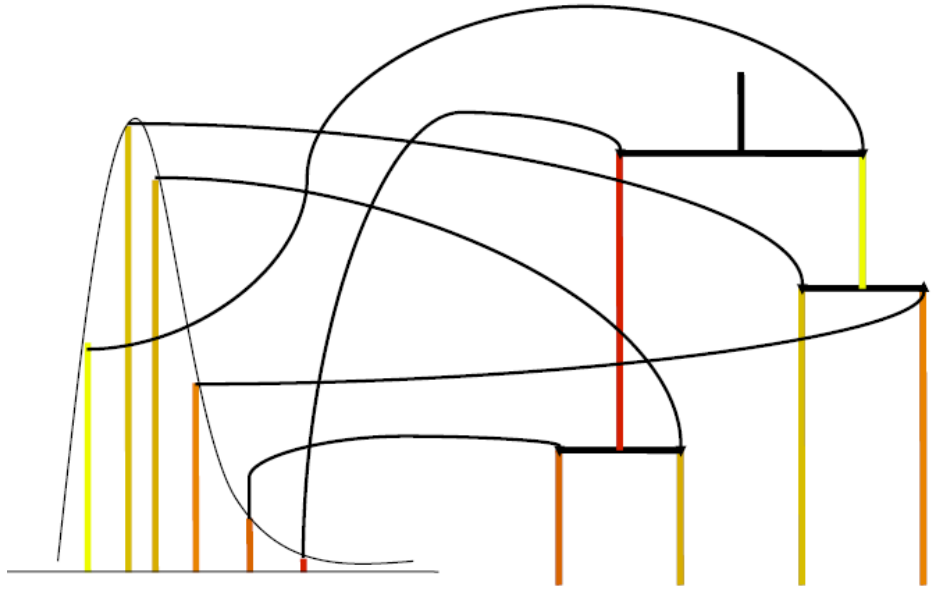
Rambaut 2016

Relaxed clock models

- Allow for among-lineage rate variation.



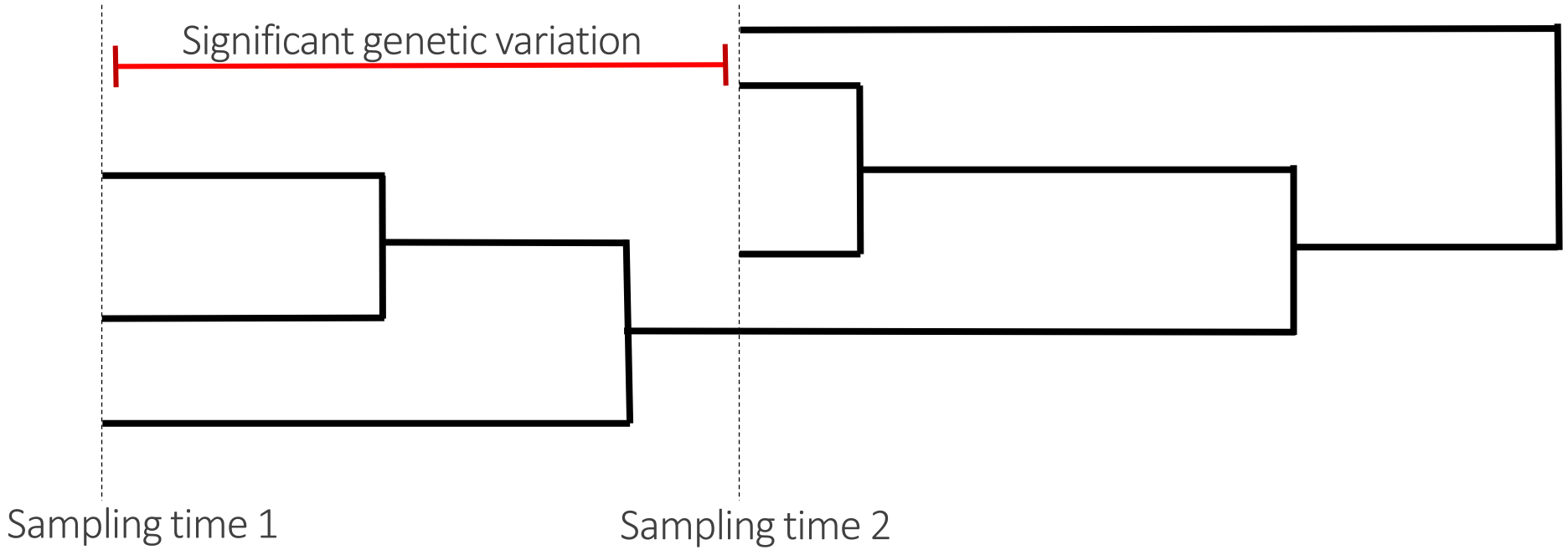
Autocorrelated clock



Uncorrelated lognormal clock

Drummond et al. 2006 Foster and Ho 2014 Rambaut 2019

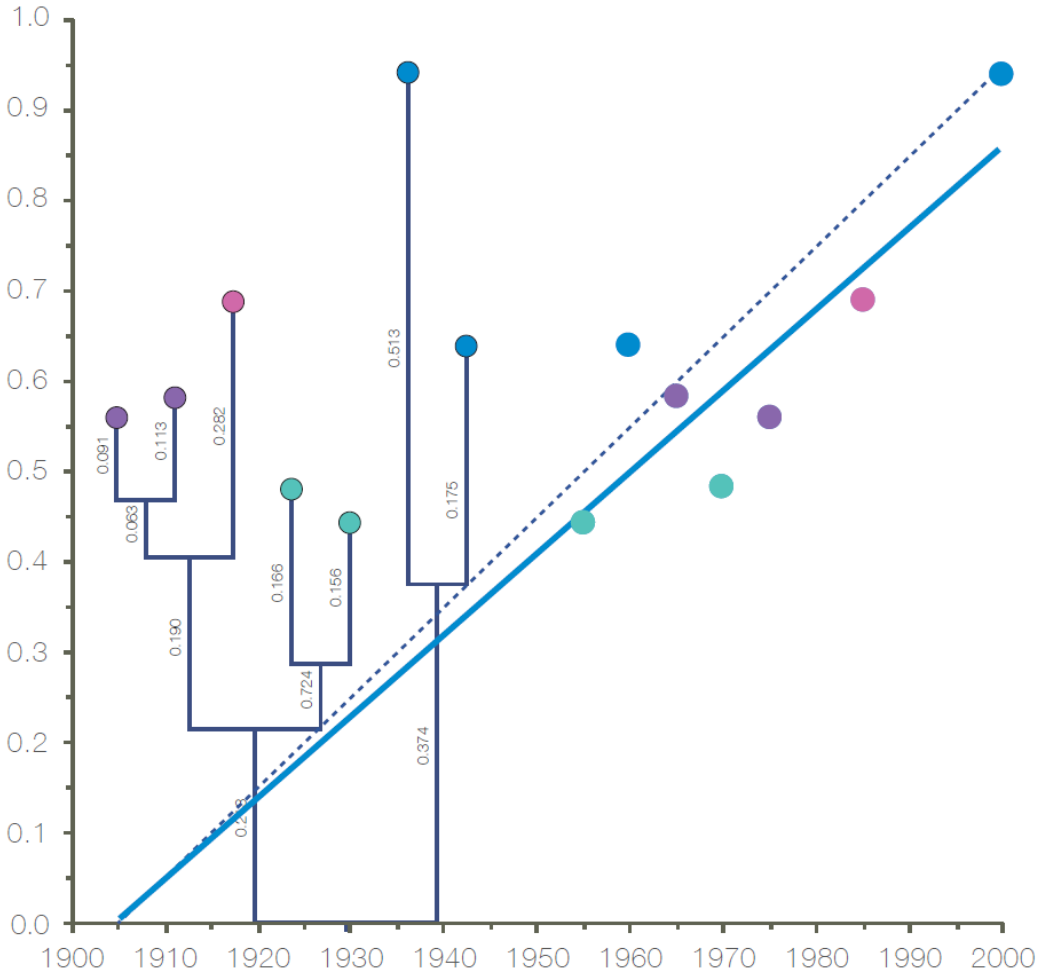
Measurable Evolving Population



Root to tip regression

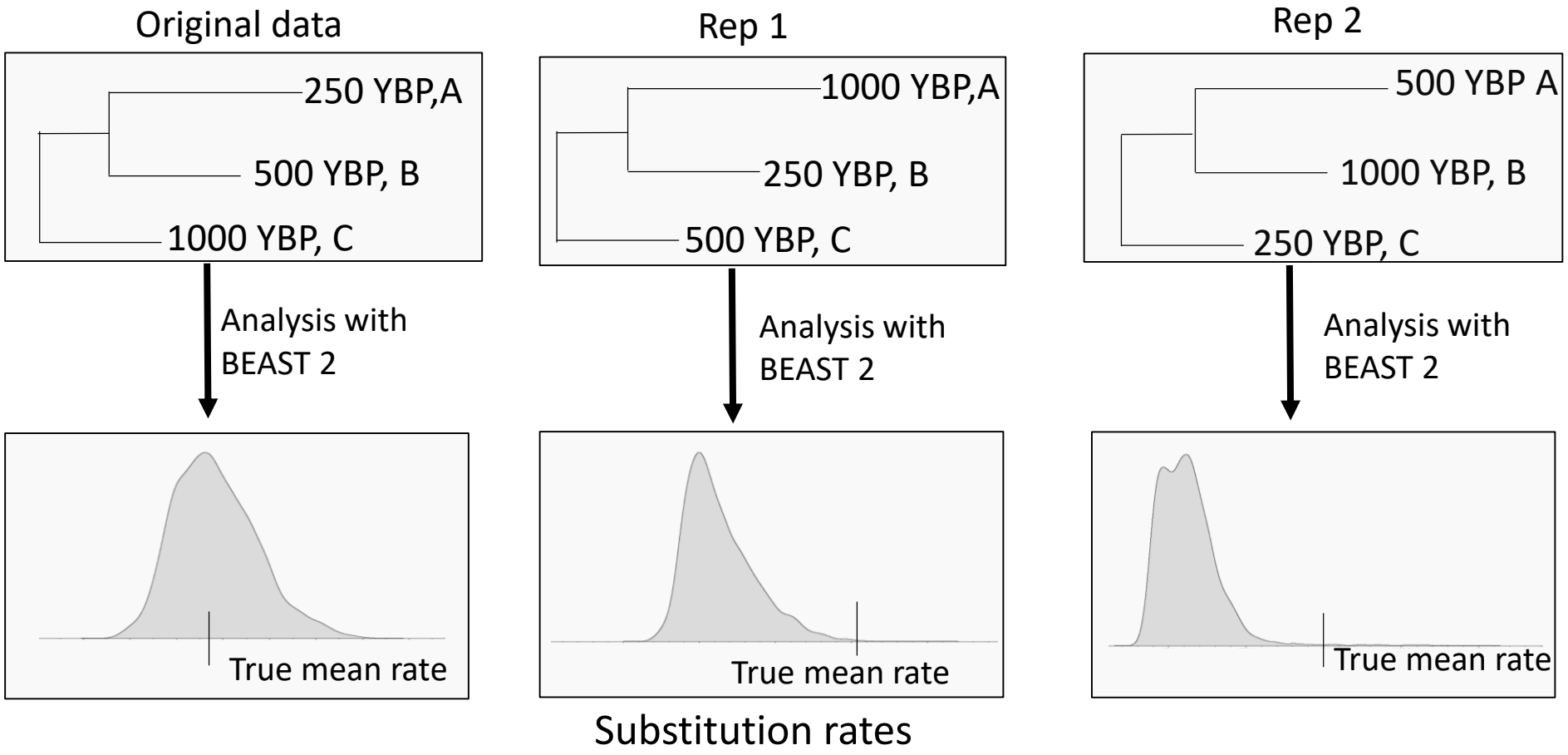
- Estimate the correlation between sampling dates and genetic distances by fitting a linear regression of root to tip genetic distances as function of time

- TempEst 



Rambaut *et al*. 2016

Date randomization test

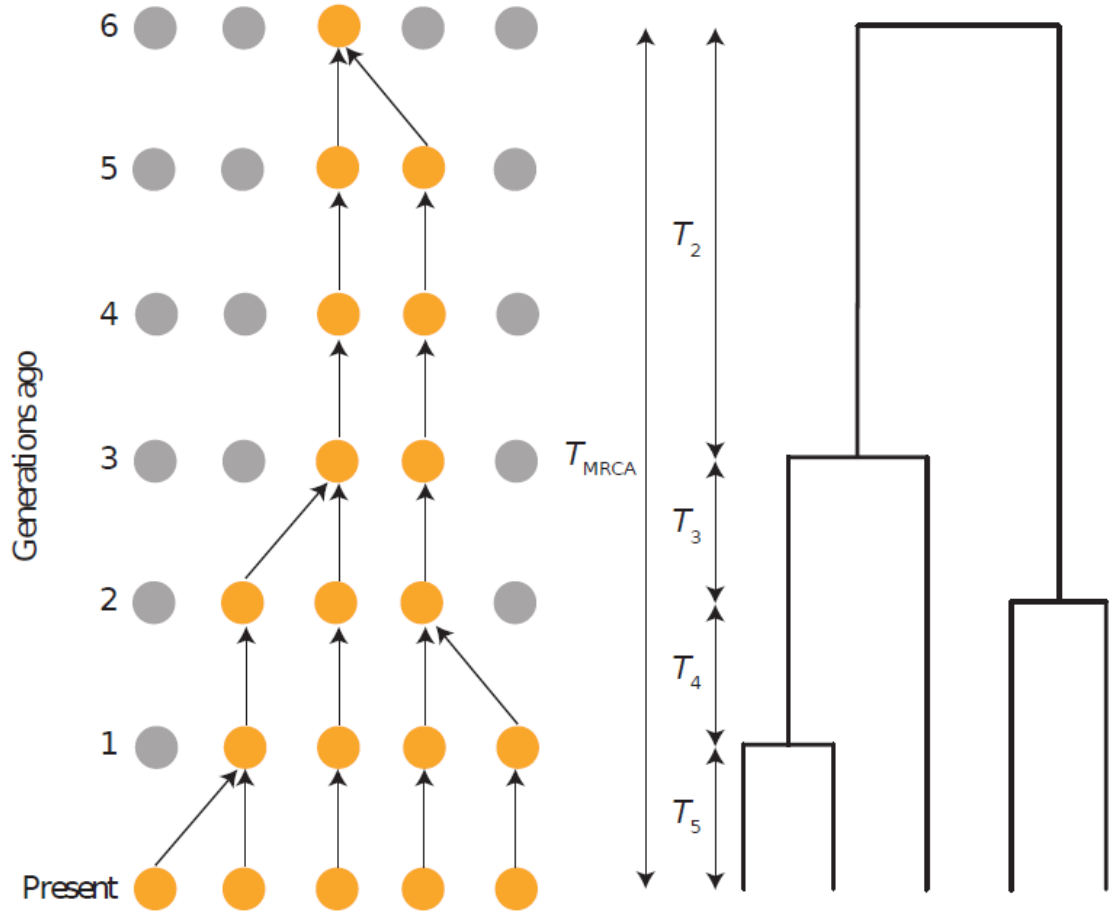


Tree Priors

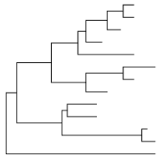
- What is the process that shaped the phylogeny?

Coalescent models

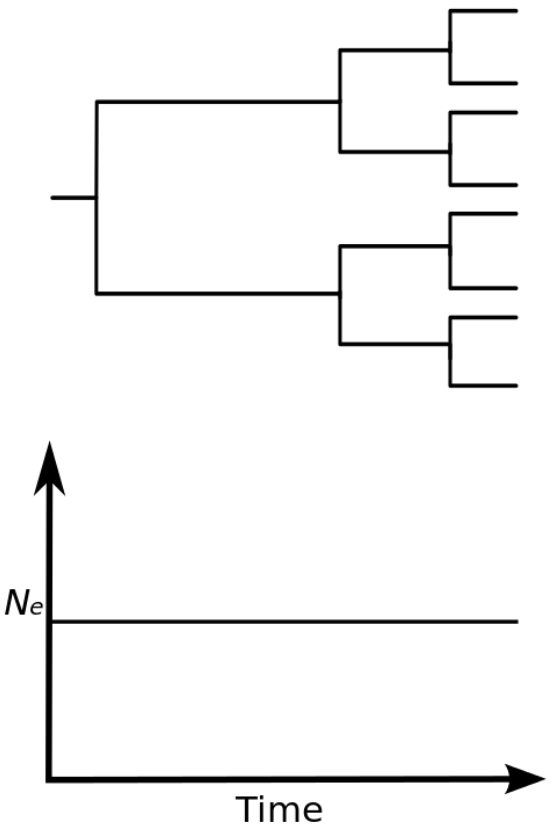
- Backwards in time
- Estimation of the effective population size N_e
- Assumes small sample from a large population



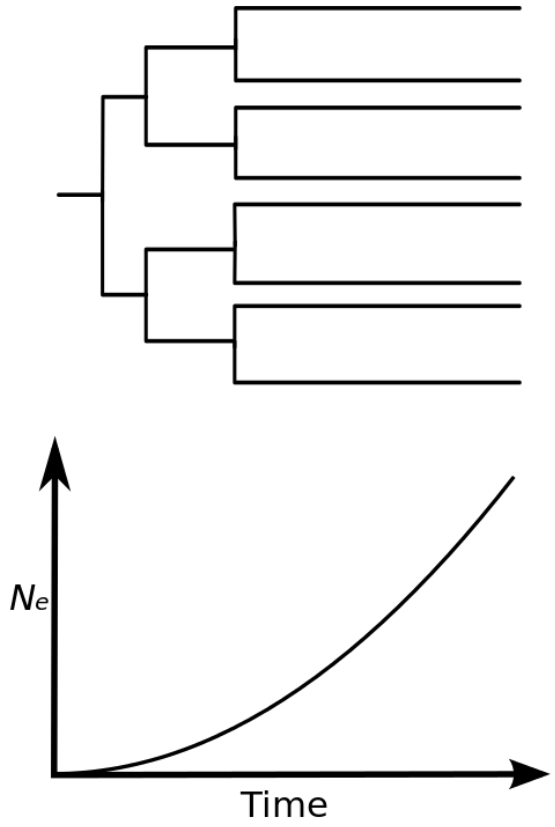
Coalescent models



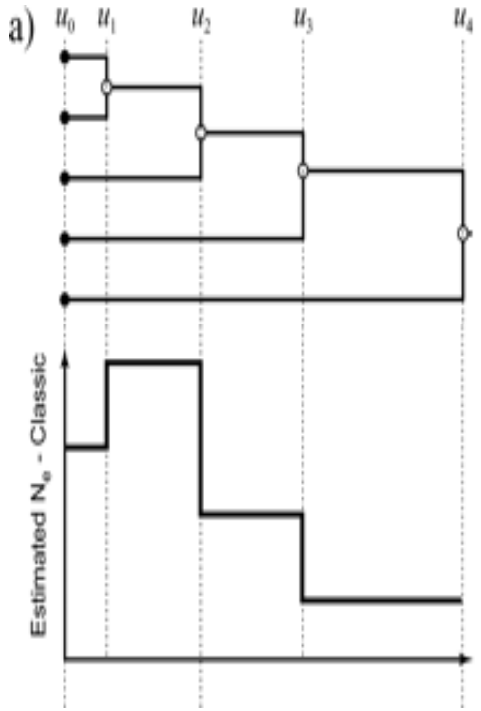
Constant Population Size



Exponential Growth

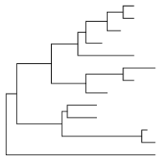


Bayesian Coalescent Skyline



Drummond *et al*. 2005 Voltz *et al*.2013

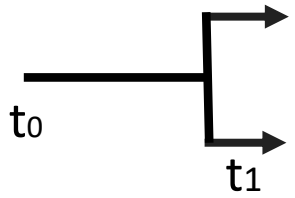
Birth-Death models



Classic Birth-death model

- Single lineage exists at time 0 in the past:

Death rate (μ)



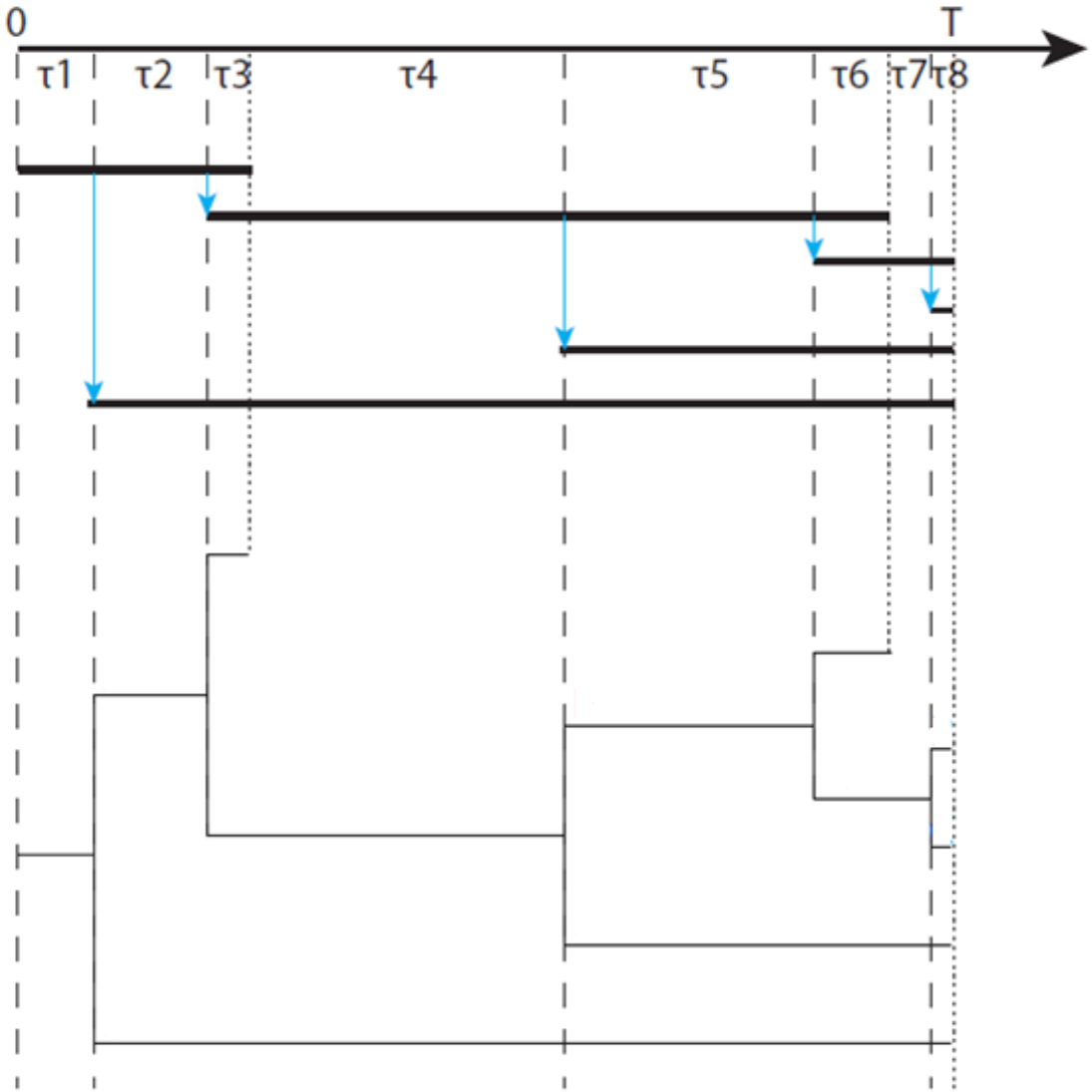
Birth rate (λ)

Yule model

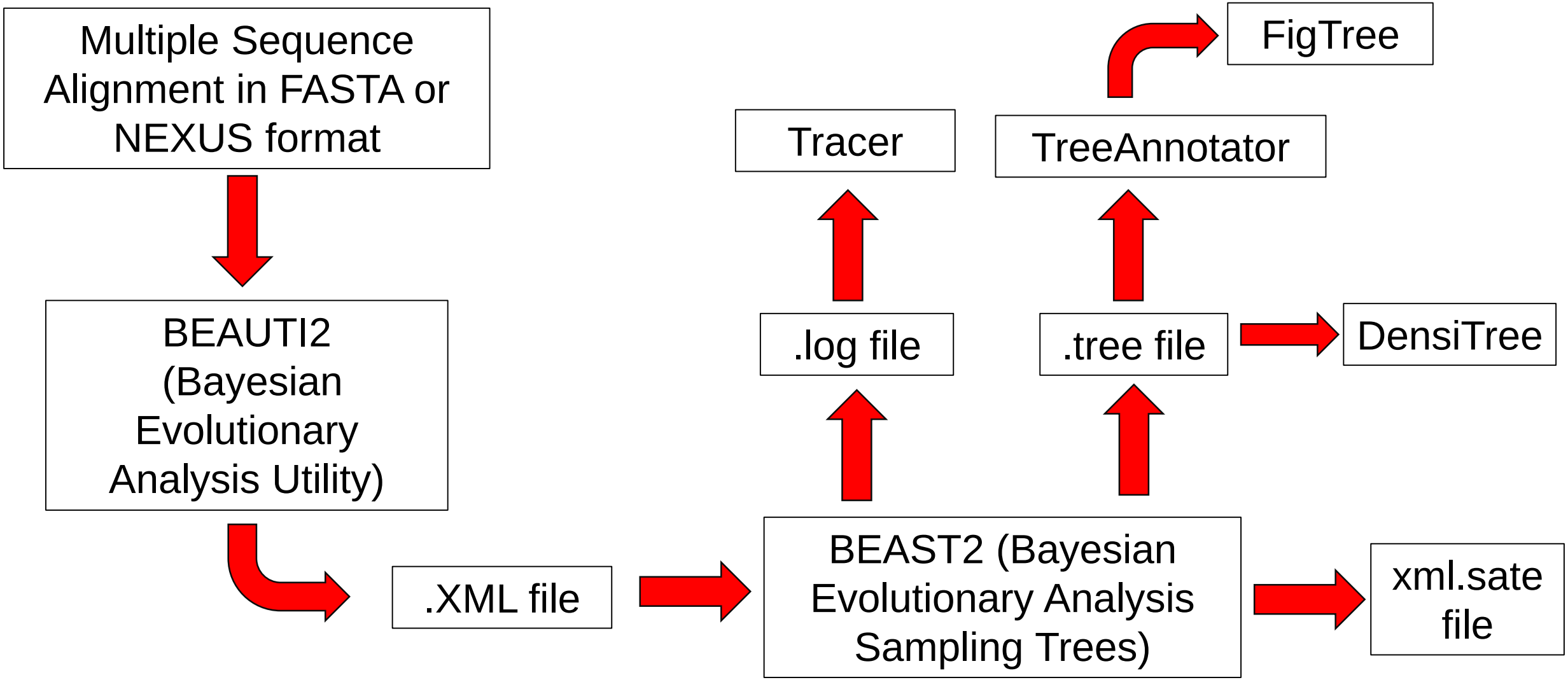
- At any given time each of the extant species are equally likely to give rise to one new species.
- Rate of speciation (λ) may vary with time.

Yule 1924 Kendall 1948

From population dynamics to phylogenetic trees



Bayesian Phylogenetic Analysis Workflow

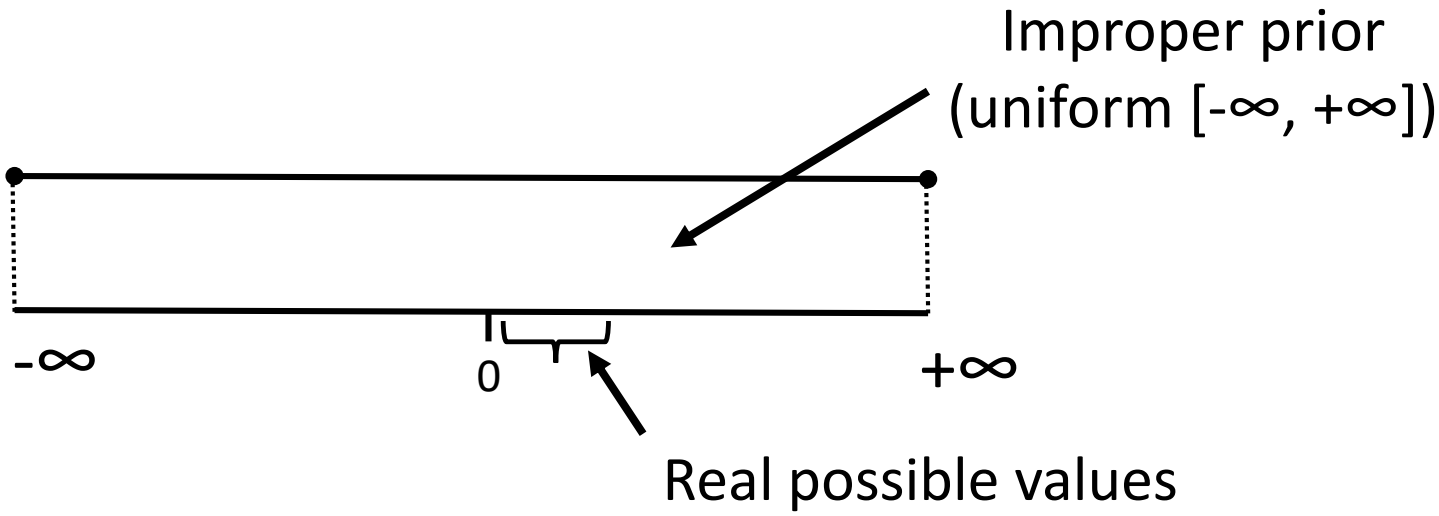


BEAUTI(2)

- Selection of model parameters.
- Settings for the tree sampling.
- Set the Priors for parameters.

➢Distribution, mean, initial value, upper an lower limits.

➢Priors should be proper (integrate to one).



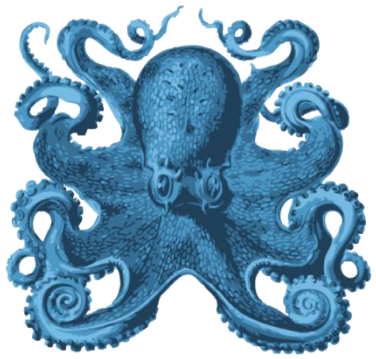
BEAST

- Bayesian phylogenetics analysis package.
- Focuses on rooted trees with time information.
- Allows estimation of:

a) Growth/decline in population

b) Dates of MRCAs

c) Rates of evolution



BEAST

Bayesian Evolutionary Analysis Sampling Trees

Drummond and Rambaut 2007

BEAST 2

- Cross-platform program for Bayesian phylogenetic analysis of molecular sequences.
- Uses a package system to perform diverse model-based analyses
- A package is developed separately to BEAST 2.

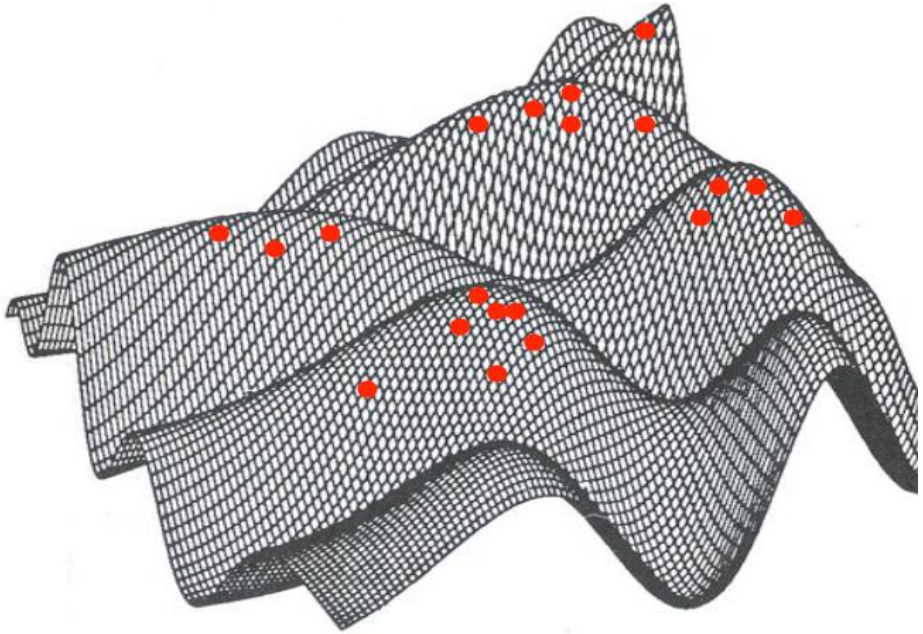
Beast2

Bayesian evolutionary analysis by sampling trees

Bouckaert *et al*. 2014

Markov Chain Monte Carlo (MCMC) Sampling

- MCMC is a sequence of random samples taken during a walk through a parameter space.
- A stochastic algorithm accepts or rejects the new state.



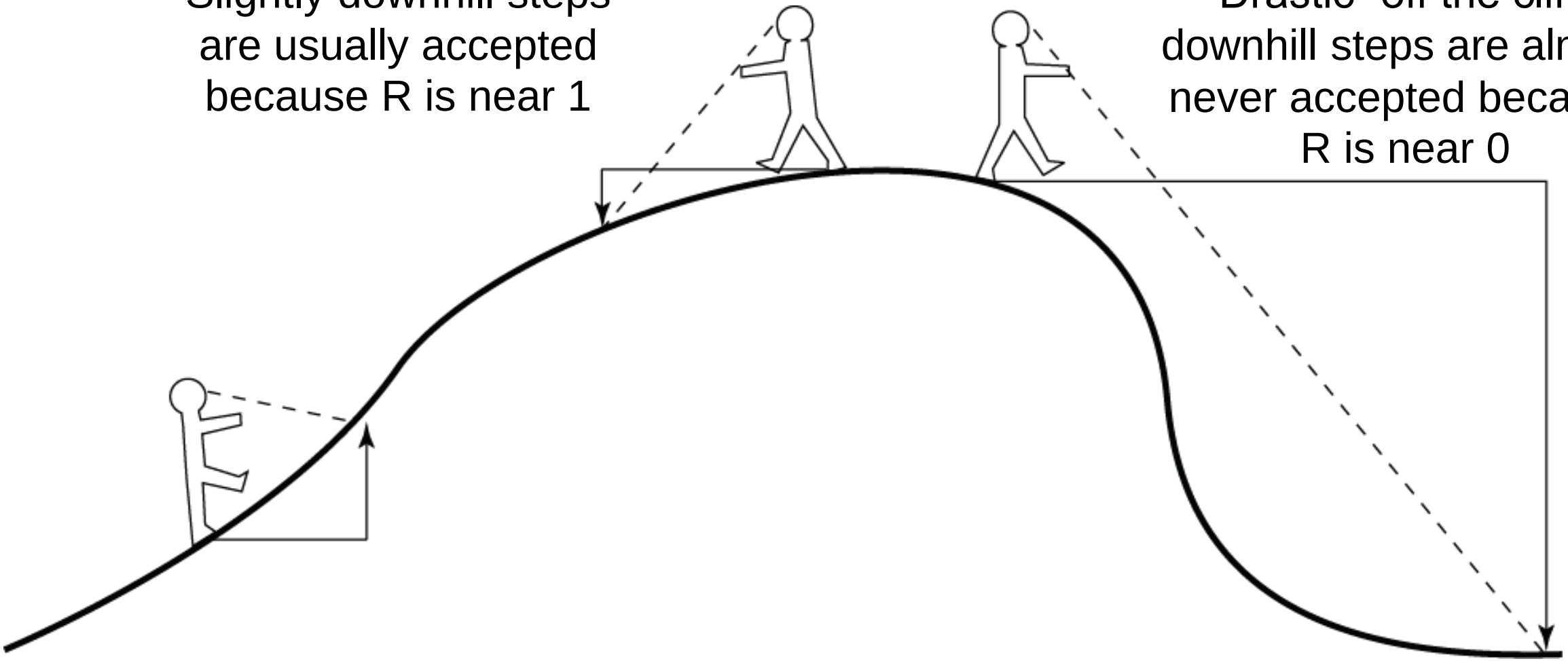
Acceptance ratio R

$$\frac{f(\tau'|X)}{f(\tau|X)} = \frac{\dfrac{f(X|\tau')\,f(\tau')}{\cancel{f(X)}}}{\dfrac{f(X|\tau)\,f(\tau)}{\cancel{f(X)}}} = \frac{f(X|\tau')\,f(\tau')}{f(X|\tau)\,f(\tau)}$$

Metropolis *et al.* 1953; Hastings 1970.

Slightly downhill steps are usually accepted because R is near 1

Drastic “off the cliff” downhill steps are almost never accepted because R is near 0



Uphill steps are always accepted because R > 1

Log file interpretation

- Tracer
- Mixing

➢Efficiency with which the MCMC algorithm samples a parameter.

➢Effective Sample Size (ESS)

a) Number of independent samples from the posterior distribution.

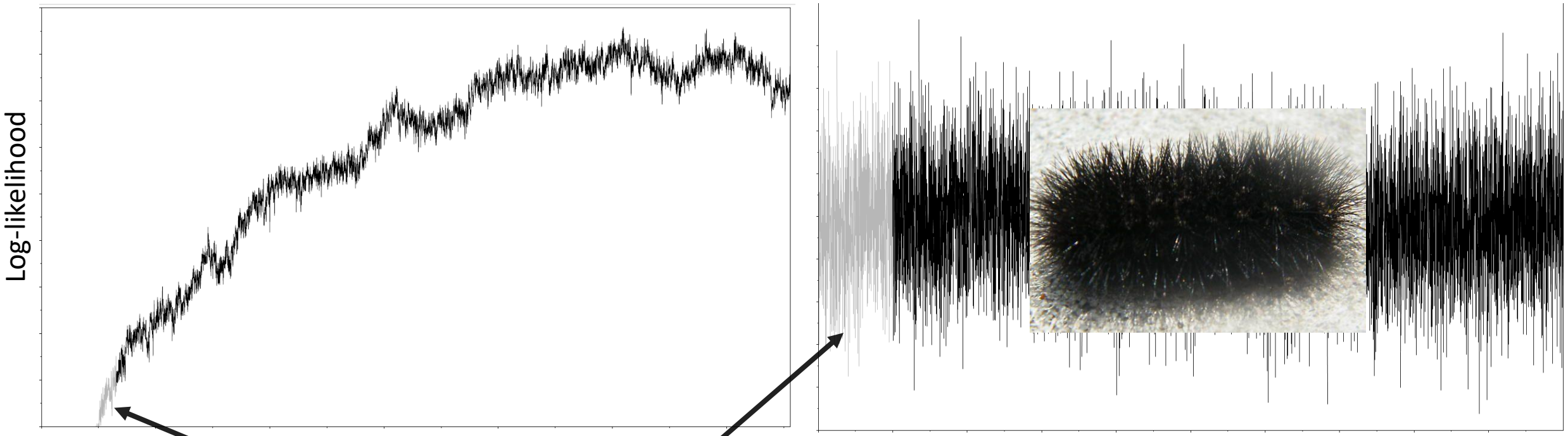
b) Number of samples divided by the Autocorrelation Time (ACT).

c) High ESS (>200)= low autocorrelation and good mixing

Rambaut *et al.* 2014

Poor mixing and convergence

Good mixing and convergence



Burn-In: Number of samples that will be discarded at the start of the run

Improve the ESS

- Increase chain length.
- Combine multiple independent runs (Tracer or LogCombiner).
- Resume runs.
- Tuning of priors and operators.

Operators

- Specify how the parameter changes as the MCMC runs.
- Scale factor: set how large a move that operator will make which will affect how often that change is accepted by the MCMC.
- Weight: specifies how often each operator is applied relative to each other.

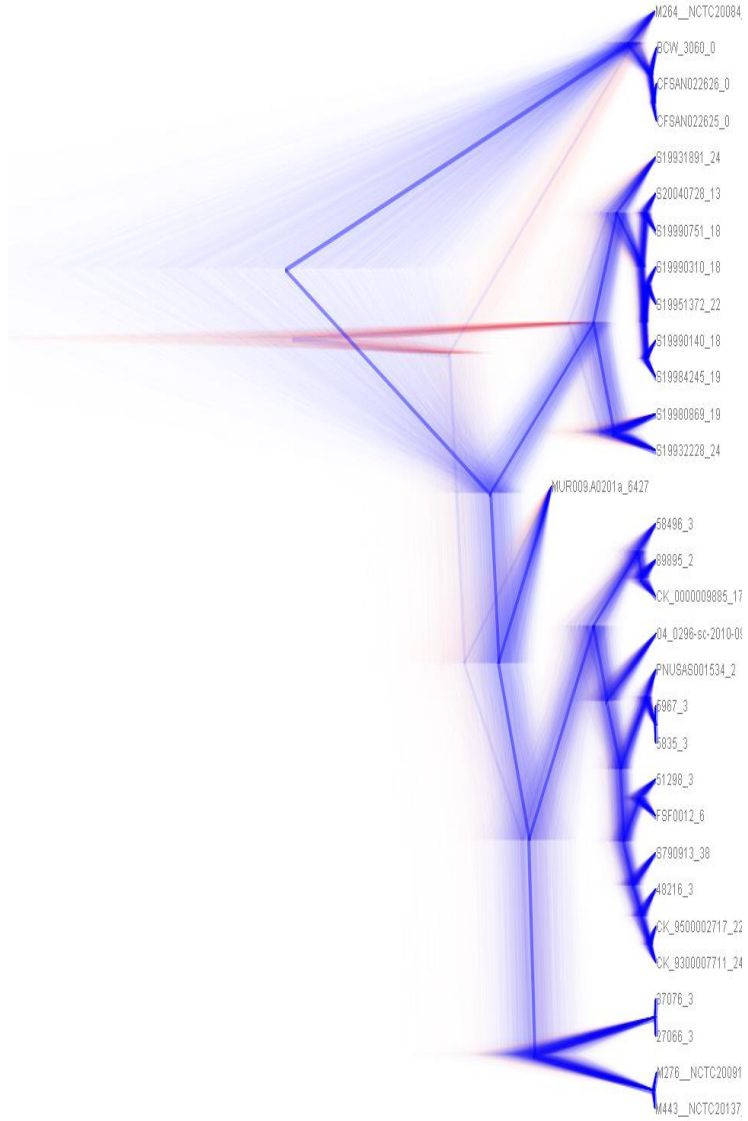
Summarize the tree file

DensiTree

- Useful to identify uncertainties in the tree topology

TreeAnnotator

- Summarizes the .trees file to produce ‘maximum clade credibility’ (MCC) trees.
- MCC: Tree with the maximum product of posterior clade probabilities.



Bouckaert and Heled 2014

Rambaut and Drummond 2012

Introduction to BEAST2 tutorial

https://taming-the-beast.org/tutorials/Introduction-to-BEAST2/