Interpreting results and summarising trees

- Refresher of MCMC traces
- Summarising estimates
  - Means, medians, mode, credible intervals, highest posterior densities
- Hypothesis support with posterior distributions
- Key traces
- Summary trees
  - Typical summary trees
  - Summarising node heights

Refresher of MCMC traces

Tracer

File Edit Analysis Help

Trace Files:

Trace File	States	Burn-In
feb15_subsample...	20000000	2000000

+ − Reload

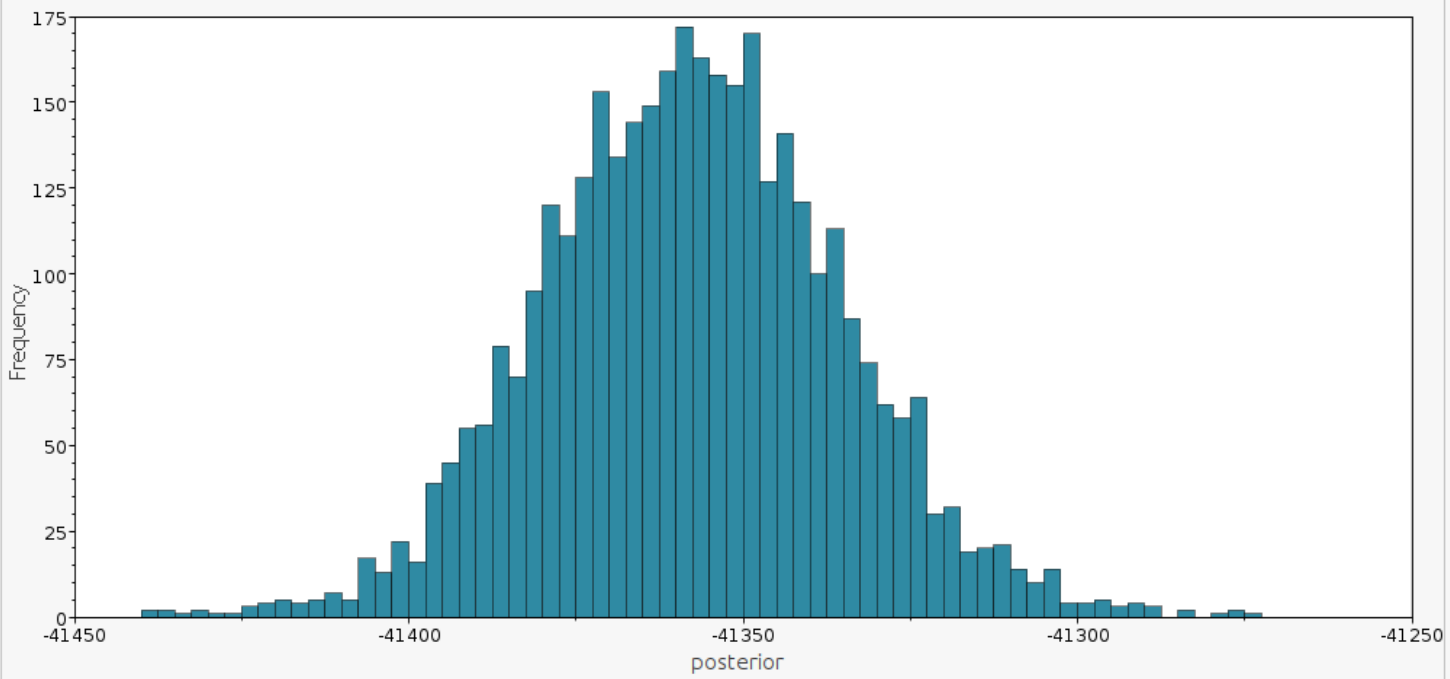
Traces:

Statistic	Mean	ESS	Ty...
posterior	-41357.667	229	R
likelihood	-41378.718	1359	R
prior	21.051	216	R
treeLikelihood	-41378.718	1359	R
TreeHeight	0.178	352	R
kappa	4.94	1753	R
gammaShape	0.47	349	R
ucldMean	1.236E-3	301	R
ucldStdev	0.272	537	R
rate.mean	1.203E-3	285	R
rate.variance	4.245E-7	679	R
rate.coefficientOfVa...	0.29	518	R
CoalescentExponent...	32.865	215	R
ePopSize	13.162	350	R
growthRate	36.71	304	R

Type: (R)eal (I)nt (C)at (T)ime * constant

Estimates | Marginal Density | Joint-Marginal | Trace

Summary Statistic	posterior
mean	-41357.667
stderr of mean	1.4729
stdev	22.2743
variance	496.143
median	-41357.8795
value range	[-41438.9032, -41273.9944]
geometric mean	n/a
95% HPD interval	[-41398.0232, -41311.276]
auto-correlation time (ACT)	78748.0028
effective sample size (ESS)	228.6
number of samples	3601



Setup... Bins: 50

Tracer

File Edit Analysis Help

Trace Files:

Trace File	States	Burn-In
feb15_subsample...	20000000	2000000

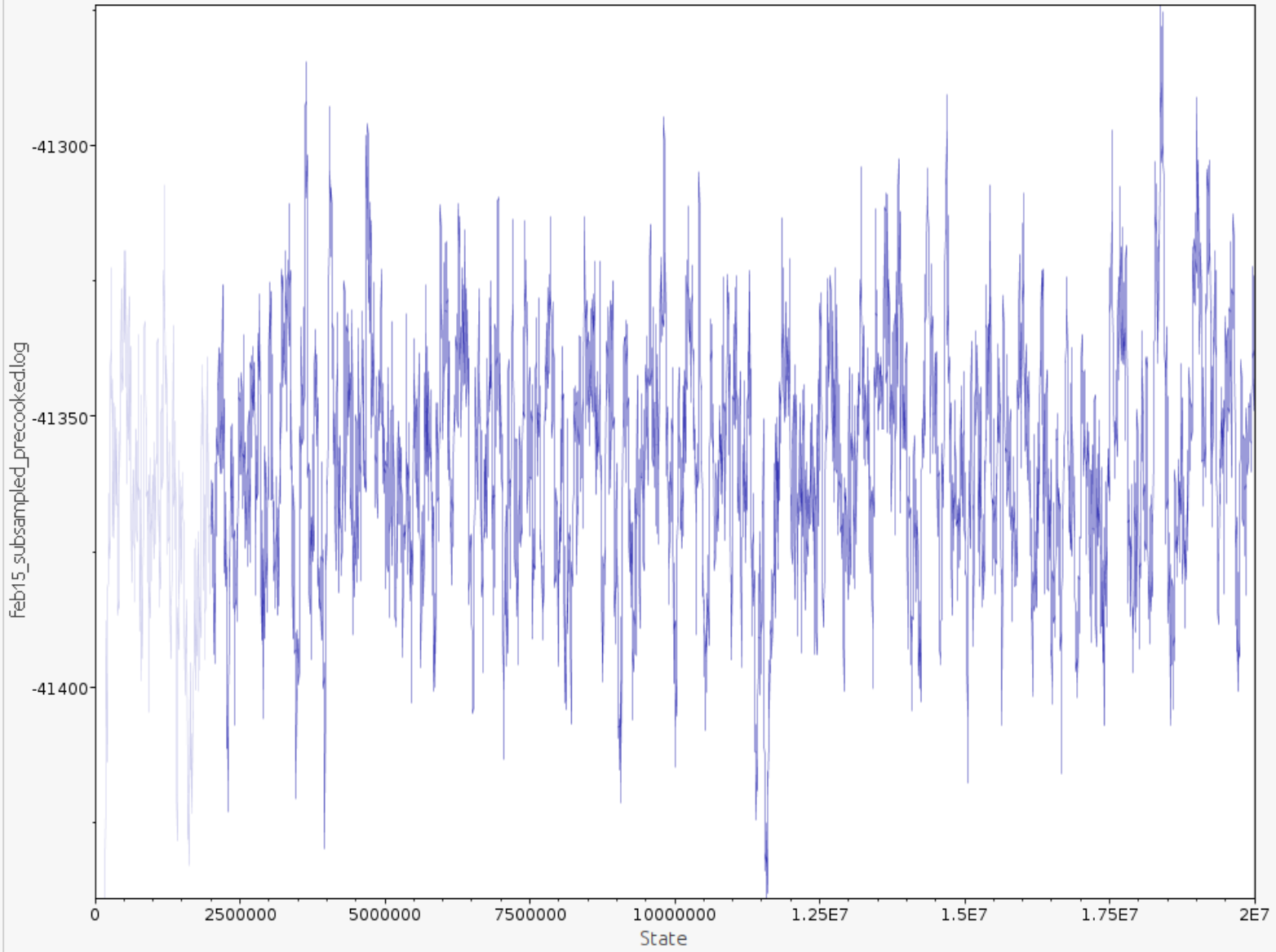
+ − Reload

Traces:

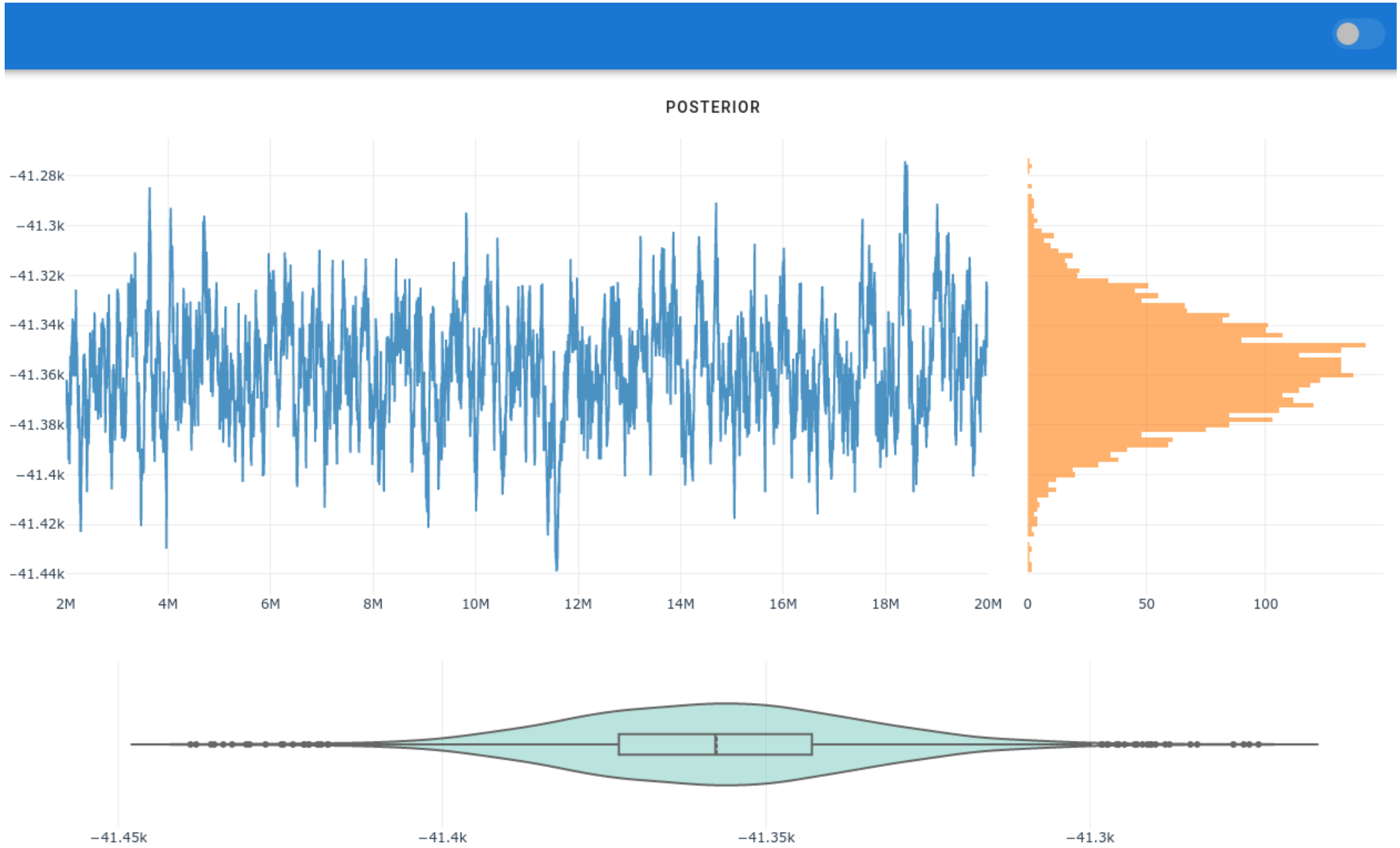
Statistic	Mean	ESS	Ty...
posterior	-41357.667	229	R
likelihood	-41378.718	1359	R
prior	21.051	216	R
treeLikelihood	-41378.718	1359	R
TreeHeight	0.178	352	R
kappa	4.94	1753	R
gammaShape	0.47	349	R
ucldMean	1.236E-3	301	R
ucldStdev	0.272	537	R
rate.mean	1.203E-3	285	R
rate.variance	4.245E-7	679	R
rate.coefficientOfVa...	0.29	518	R
CoalescentExponent...	32.865	215	R
ePopSize	13.162	350	R
growthRate	36.71	304	R

Type: (R)eal (I)nt (C)at (T)ime * constant

Estimates | Marginal Density | Joint-Marginal | Trace



Show Burn-in | Sample only | Draw line plot | Legend: None | Colour by: Trace


POSTERIOR
−41.28k
−41.3k
−41.32k
−41.34k
−41.36k
−41.38k
−41.4k
−41.42k
−41.44k
2M
4M
6M
8M
10M
12M
14M
16M
18M
20M
0
50
100
−41.45k
−41.4k
−41.35k
−41.3k

Tracer

File Edit Analysis Help

Trace Files:

Trace File	States	Burn-In
feb15_subsample...	20000000	2000000

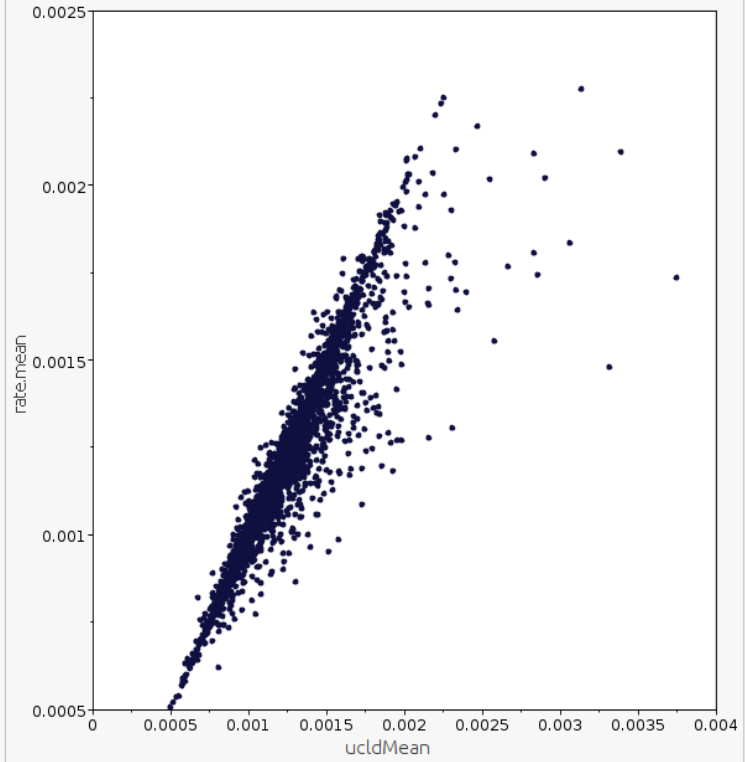
+ − Reload

Traces:

Statistic	Mean	ESS	Ty...
posterior	-41357.667	229	R
likelihood	-41378.718	1359	R
prior	21.051	216	R
treeLikelihood	-41378.718	1359	R
TreeHeight	0.178	352	R
kappa	4.94	1753	R
gammaShape	0.47	349	R
ucldMean	1.236E-3	301	R
ucldStdev	0.272	537	R
rate.mean	1.203E-3	285	R
rate.variance	4.245E-7	679	R
rate.coefficientOfVa...	0.29	518	R
CoalescentExponent...	32.865	215	R
ePopSize	13.162	350	R
growthRate	36.71	304	R

Type: (R)eal (I)nt (C)at (T)ime * constant

Estimates μ | Marginal Density | Joint-Marginal | Trace



Setup... Sample only Draw as points Use translucency

Tracer

File Edit Analysis Help

Trace Files:

Trace File	States	Burn-In
feb15_subsample...	20000000	2000000

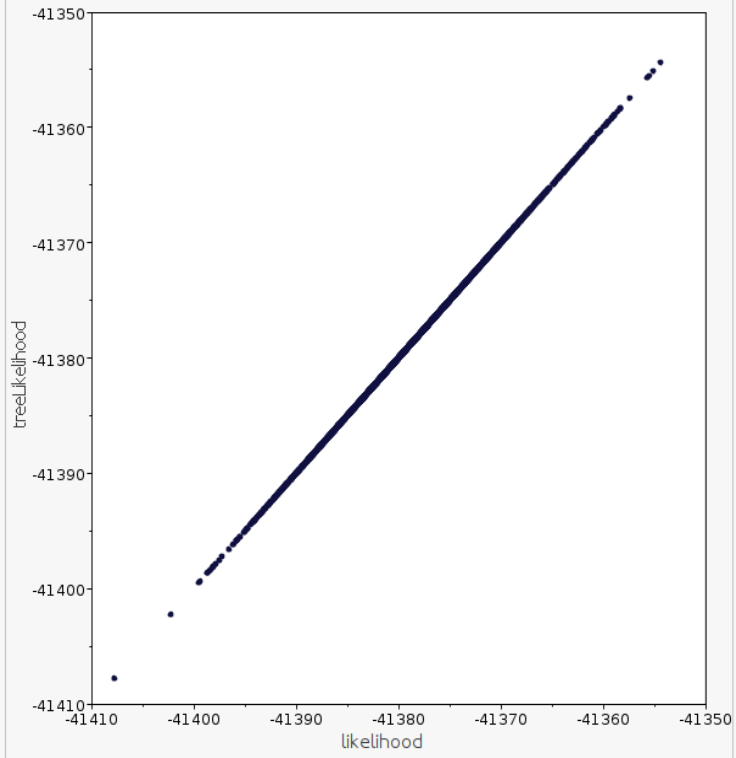
+ − Reload

Traces:

Statistic	Mean	ESS	Ty...
posterior	-41357.667	229	R
likelihood	-41378.718	1359	R
prior	21.051	216	R
treeLikelihood	-41378.718	1359	R
TreeHeight	0.178	352	R
kappa	4.94	1753	R
gammaShape	0.47	349	R
ucldMean	1.236E-3	301	R
ucldStdev	0.272	537	R
rate.mean	1.203E-3	285	R
rate.variance	4.245E-7	679	R
rate.coefficientOfVa...	0.29	518	R
CoalescentExponent...	32.865	215	R
ePopSize	13.162	350	R
growthRate	36.71	304	R

Type: (R)eal (I)nt (C)at (T)ime * constant

Estimates μ | Marginal Density | Joint-Marginal | Trace



Setup... Sample only Draw as points Use translucency

Tracer

File Edit Analysis Help

Trace Files:

Trace File	States	Burn-In
feb15_subsample...	20000000	2000000

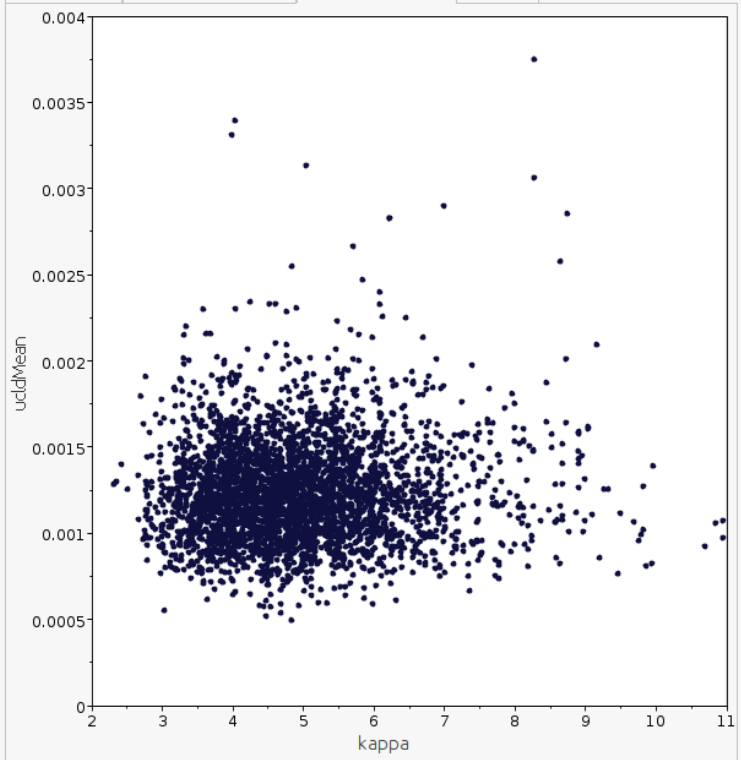
+ − Reload

Traces:

Statistic	Mean	ESS	Ty...
posterior	-41357.667	229	R
likelihood	-41378.718	1359	R
prior	21.051	216	R
treeLikelihood	-41378.718	1359	R
TreeHeight	0.178	352	R
kappa	4.94	1753	R
gammaShape	0.47	349	R
ucldMean	1.236E-3	301	R
ucldStdev	0.272	537	R
rate.mean	1.203E-3	285	R
rate.variance	4.245E-7	679	R
rate.coefficientOfVa...	0.29	518	R
CoalescentExponent...	32.865	215	R
ePopSize	13.162	350	R
growthRate	36.71	304	R

Type: (R)eal (I)nt (C)at (T)ime * constant

Estimates π Marginal Density Joint-Marginal Trace



Setup... Sample only Draw as points Use translucency

Tracer

File Edit Analysis Help

Trace Files:

Trace File	States	Burn-In
feb15_subsample...	20000000	2000000

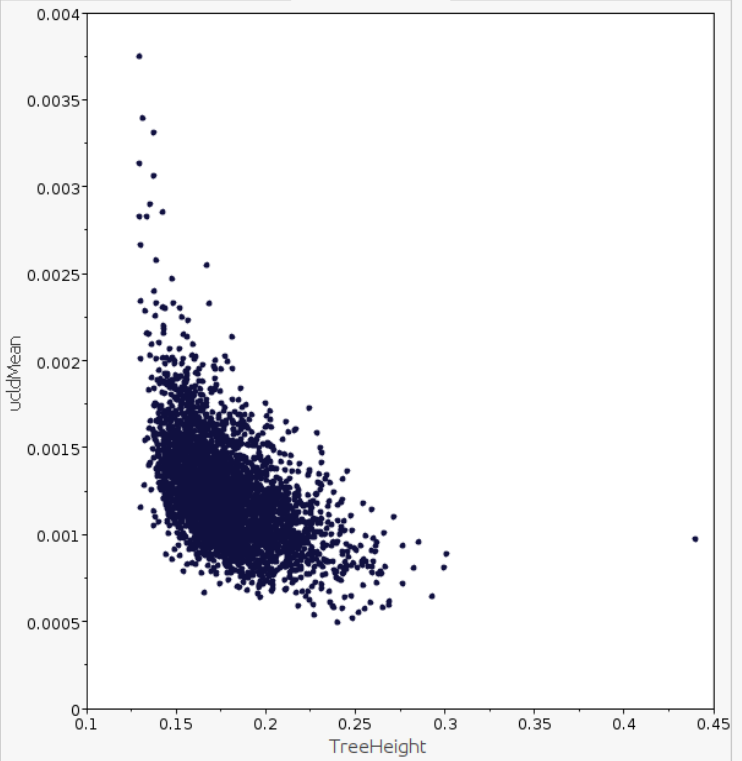
+ − Reload

Traces:

Statistic	Mean	ESS	Ty...
posterior	-41357.667	229	R
likelihood	-41378.718	1359	R
prior	21.051	216	R
treeLikelihood	-41378.718	1359	R
TreeHeight	0.178	352	R
kappa	4.94	1753	R
gammaShape	0.47	349	R
ucldMean	1.236E-3	301	R
ucldStdev	0.272	537	R
rate.mean	1.203E-3	285	R
rate.variance	4.245E-7	679	R
rate.coefficientOfVa...	0.29	518	R
CoalescentExponent...	32.865	215	R
ePopSize	13.162	350	R
growthRate	36.71	304	R

Type: (R)eal (I)nt (C)at (T)ime * constant

Estimates π Marginal Density Joint-Marginal Trace



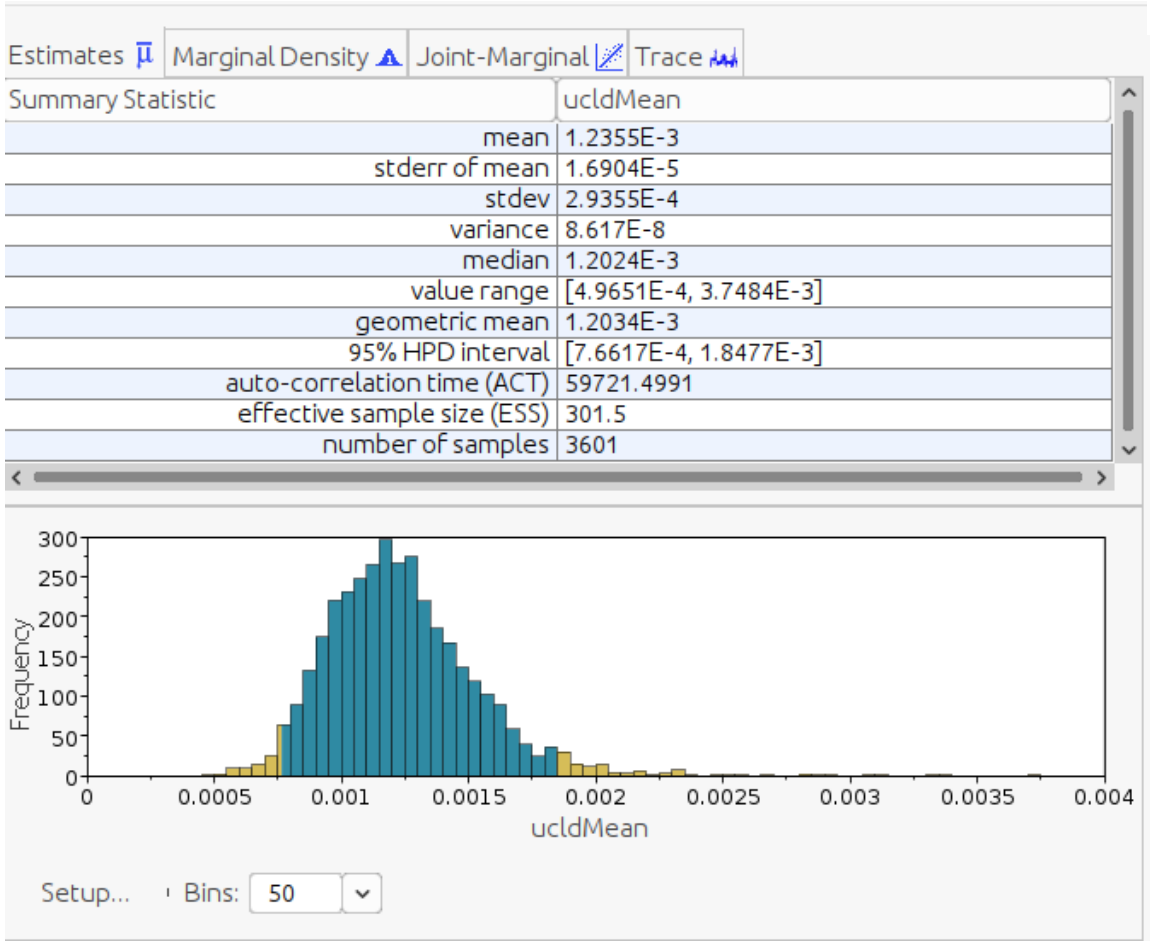
Setup... Sample only Draw as points Use translucency

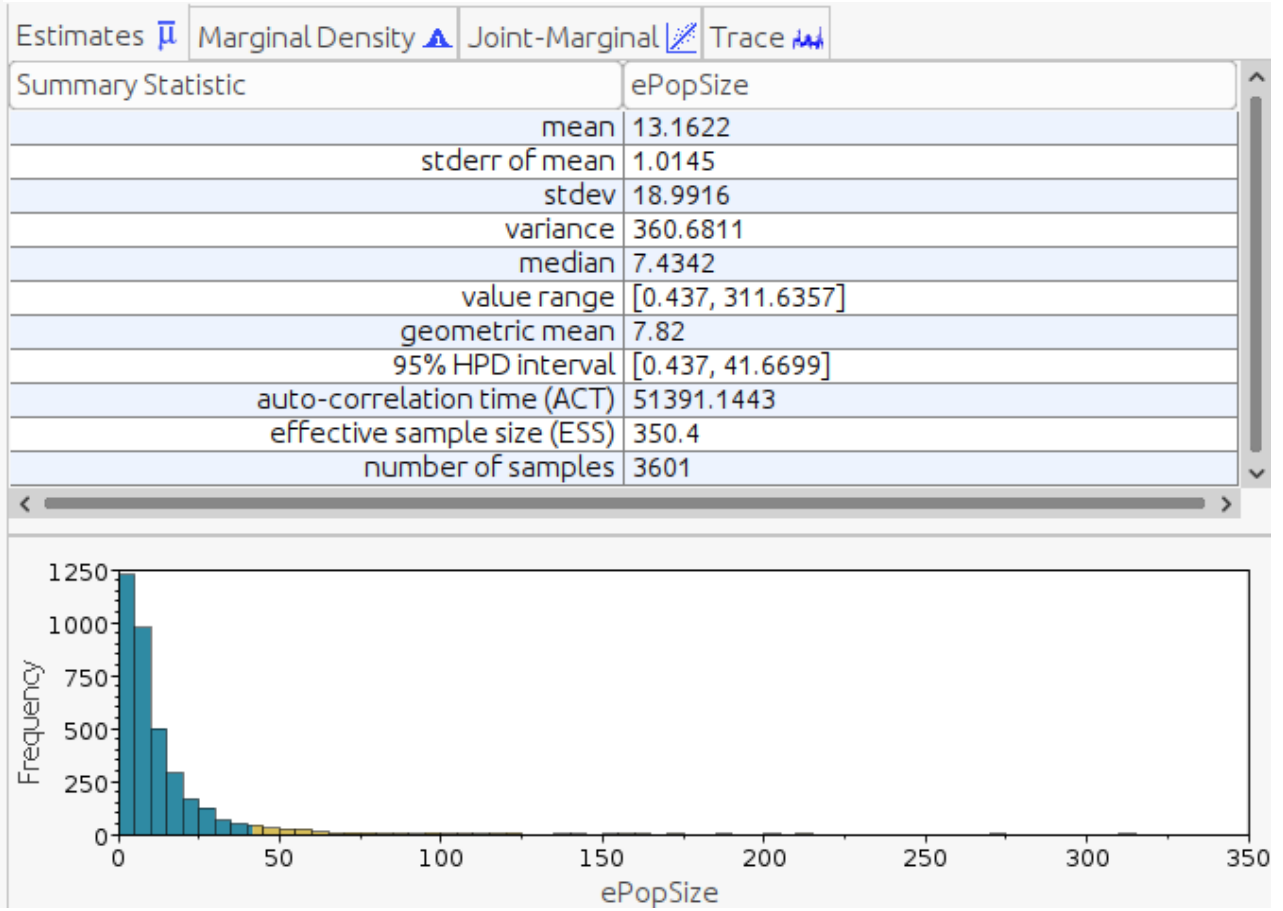
Concept summary

- MCMC sampling frequency is proportional to the posterior probability.

- Correlation between parameters is sometimes expected (but not for most parameters -> overparameterisation).

Summarising estimates

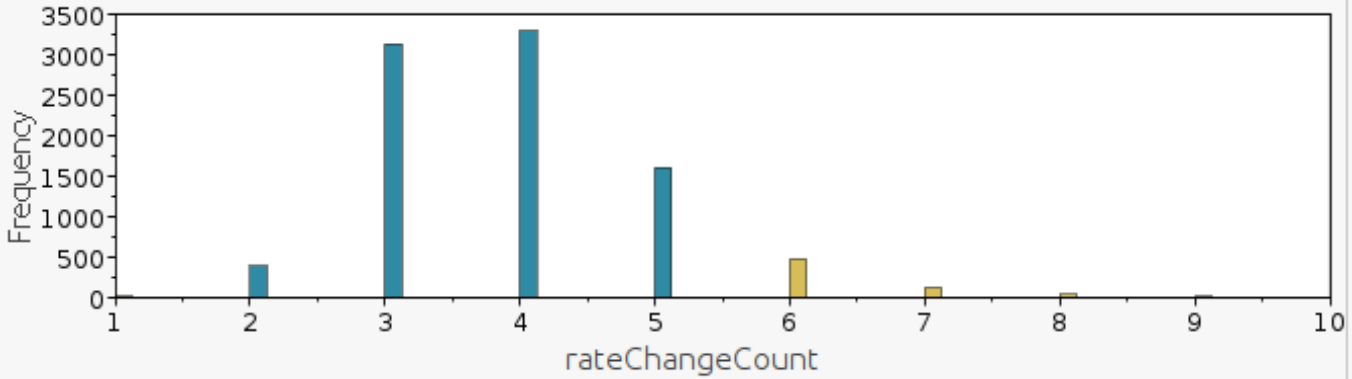




- Choice of one summary statistic (e.g. mean, median) depends on the shape of the distribution.

Estimates | Marginal Density | Joint-Marginal | Trace

Summary Statistic	rateChangeCount
mean	3.9008
stderr of mean	0.0286
stdev	1.0327
variance	1.0665
median	4
value range	[1, 9]
geometric mean	3.769
95% HPD interval	[2, 6]
auto-correlation time (ACT)	34568.7283
effective sample size (ESS)	1301.9
number of samples	9001



- For discrete statistics we can use the mode=maximum a posteriori estimate (MAP).

Reporting uncertainty

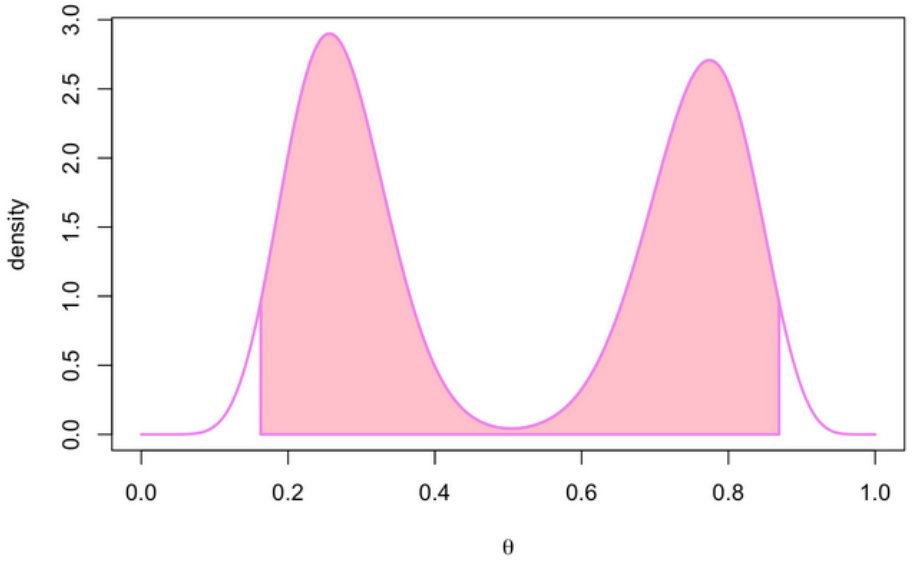
- Credible interval: 'given our data there is a 0.95 probability that our parameter falls in a range.'

- Highest posterior density: 'every point within the interval has a higher density than any point outside.'

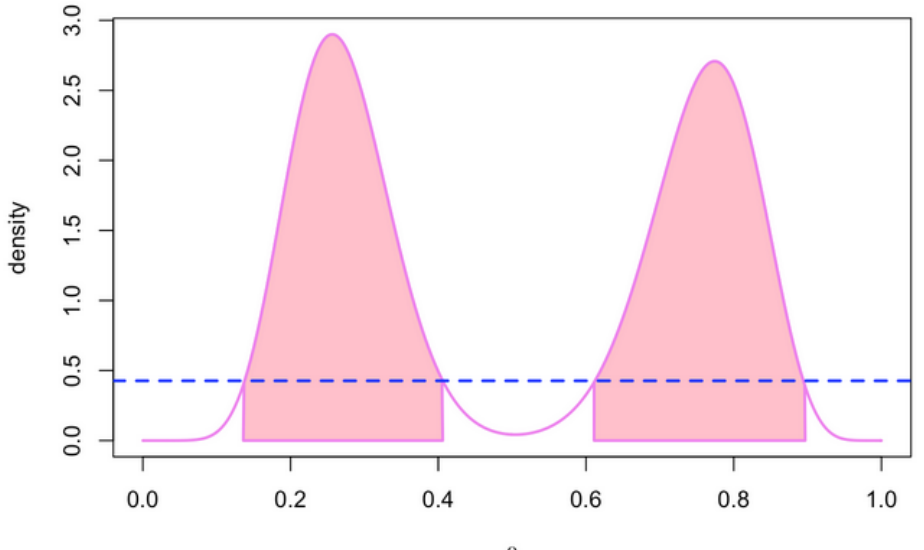
- Confidence interval: 'there is a 0.95 probability that if we repeat this experiment 100 times, the confidence interval will include the true value 95 times.'

Reporting uncertainty

Credible intervals are very similar to HPDs when the distribution is unimodal. Can be calculated with quantiles or percentiles.



HPDs are very useful for bimodal distributions or those that are very skewed. They require additional calculations (in R use package HDinterval.



Figs from vioshyvo.github.io

Concept summary

- All parameters and statistics have uncertainties.
  - Report using credible intervals or highest posterior densities NOT confidence intervals.
- The choice of summary values (e.g. mean) depends on the distribution.

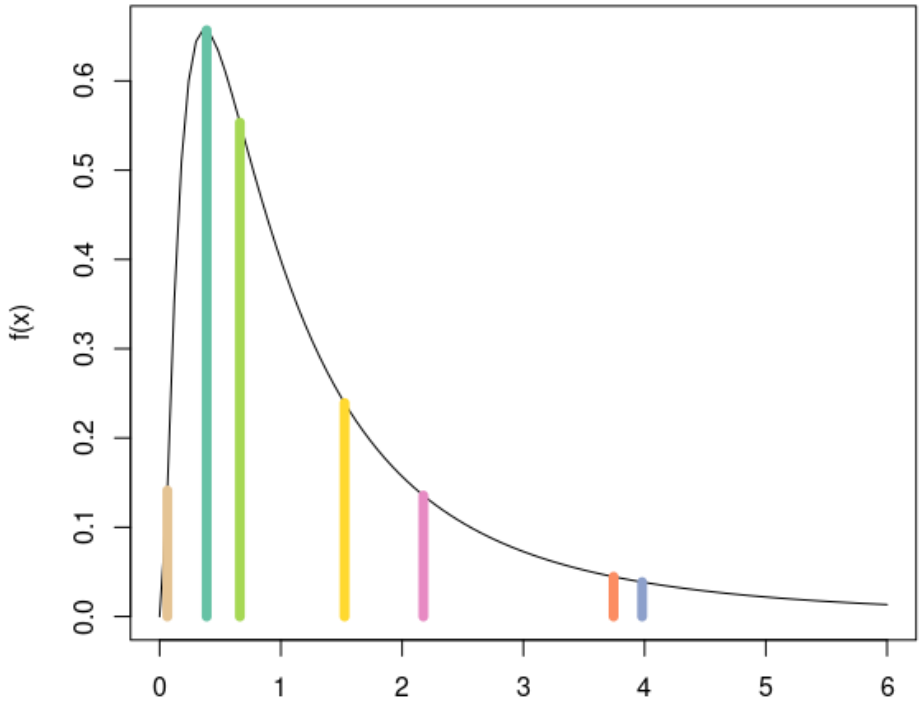
- Reporting the MAP value is an alternative to the mean, particularly if the distribution is not very skewed or if it is discrete.

Hypothesis support using posterior distributions

Bayesian hypothesis testing

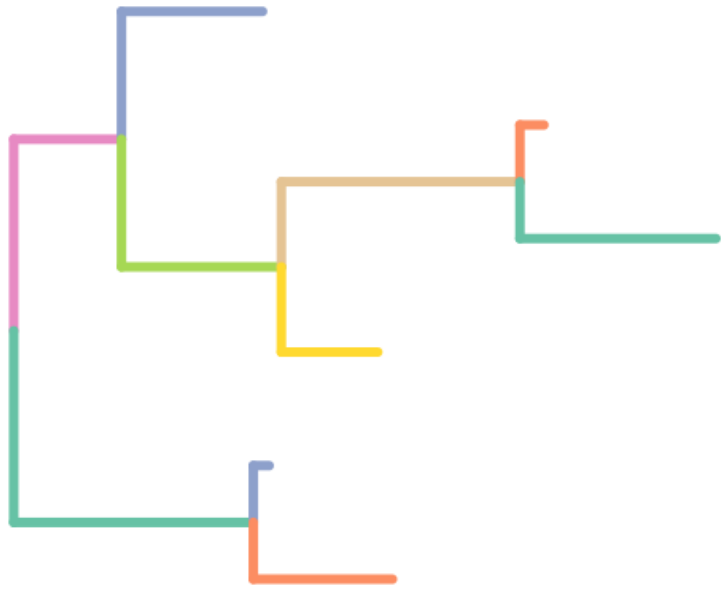
- 'Formal' methods via Bayes factors and marginal likelihoods
  - Requires additional computation, proper priors, and careful prior selection.
- Inspection of the posterior distributions
  - Straight-forward if model is correctly parameterised.
  - May require setting up complex hierarchical models.
  - Only for 'nested' models
- Posterior model checking
  - Requires simulations and data summaries.
  - Can assess the overall, not just relative, fit of models.
- Bayesian model averaging
  - Alternative to marginal likelihood calculations.
  - Traversing model space can be difficult.

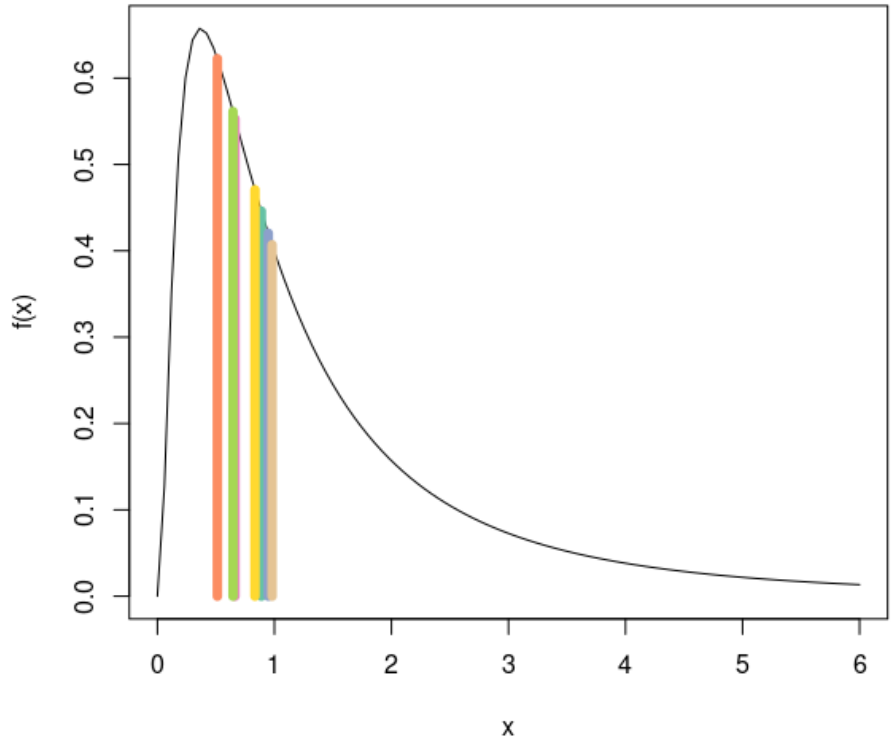
We still have the responsibility of proposing a pool of models!



The mean of the lognormal distribution (**ucld.mean**) is 0.1 and the sd (**ucld.stdev**) is 1.
The mean of the branch rates (*meanRate*) is 1.79 and the *sd* is 1.58

Coefficient of rate variation = sd of branch rates / meanRate
= 0.88

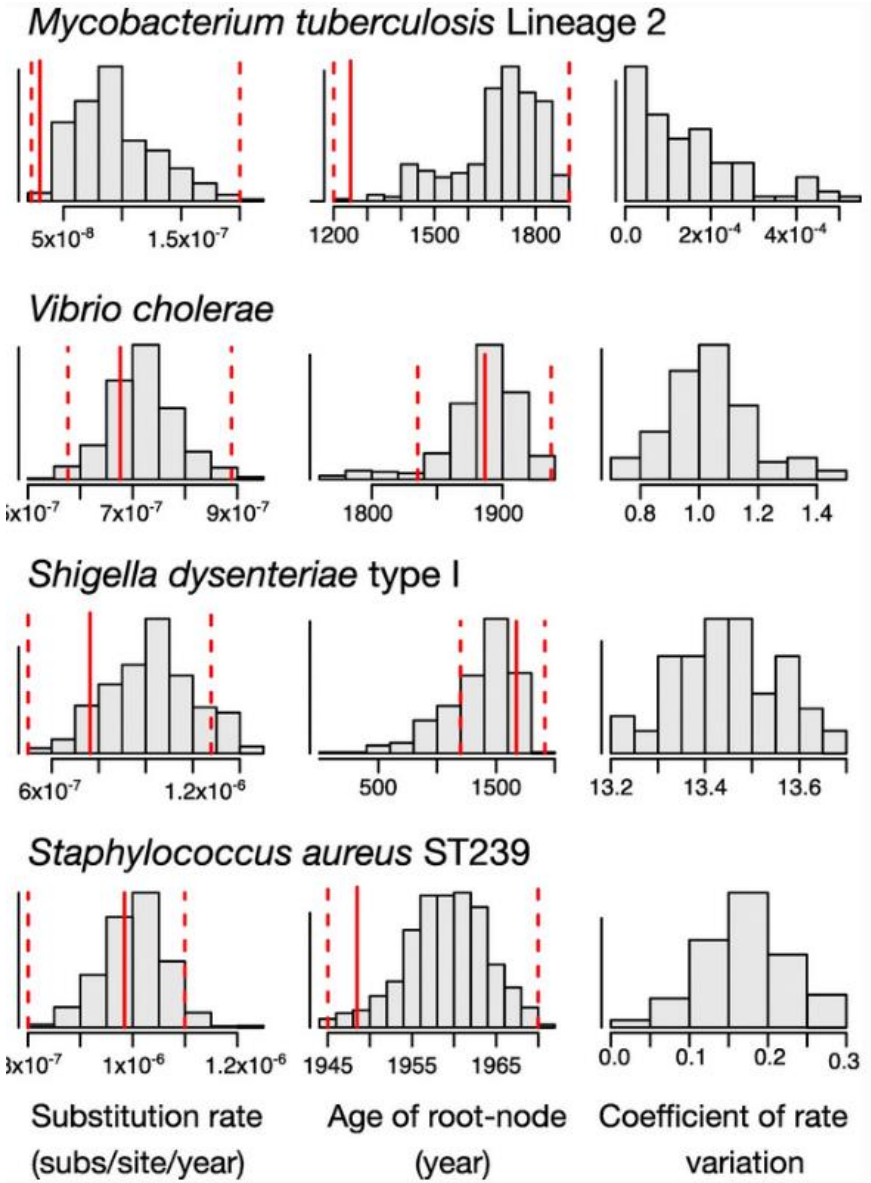


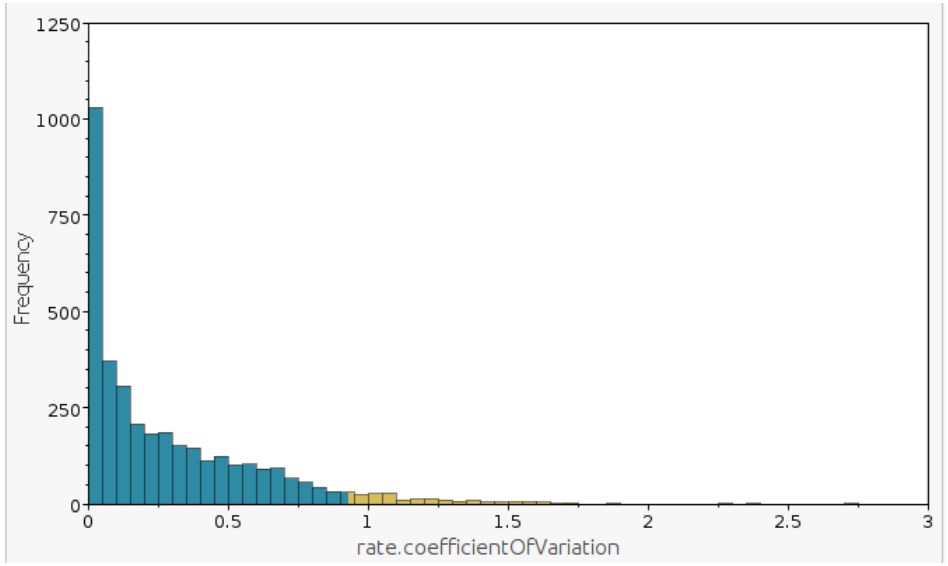


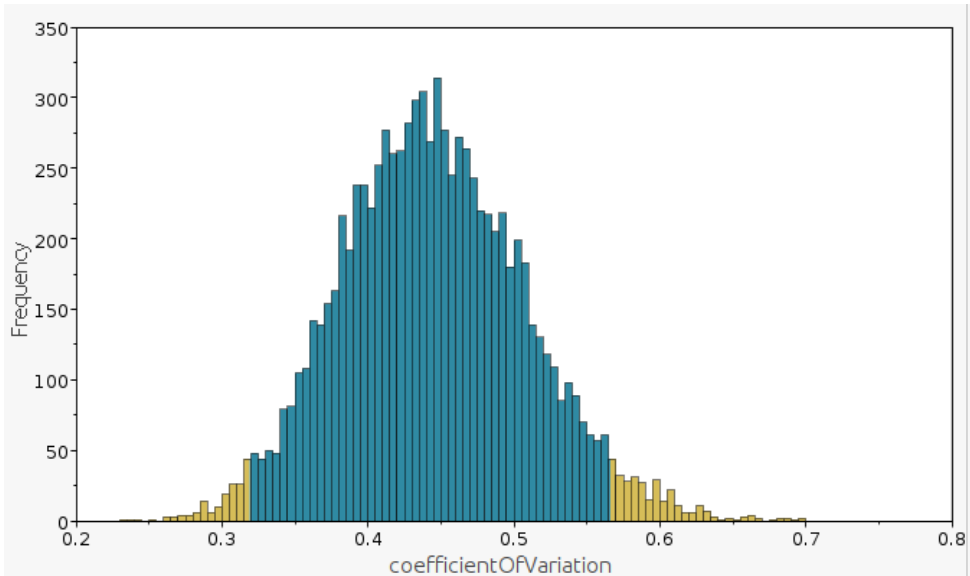
The mean of the lognormal distribution (**ucld.mean**) is 0.1 and the sd (**ucld.stdev**) is 1.
The mean of the branch rates (*meanRate*) is 0.78 and the *sd* is 0.17

Coefficient of rate variation = sd of branch rates / meanRate
= 0.22

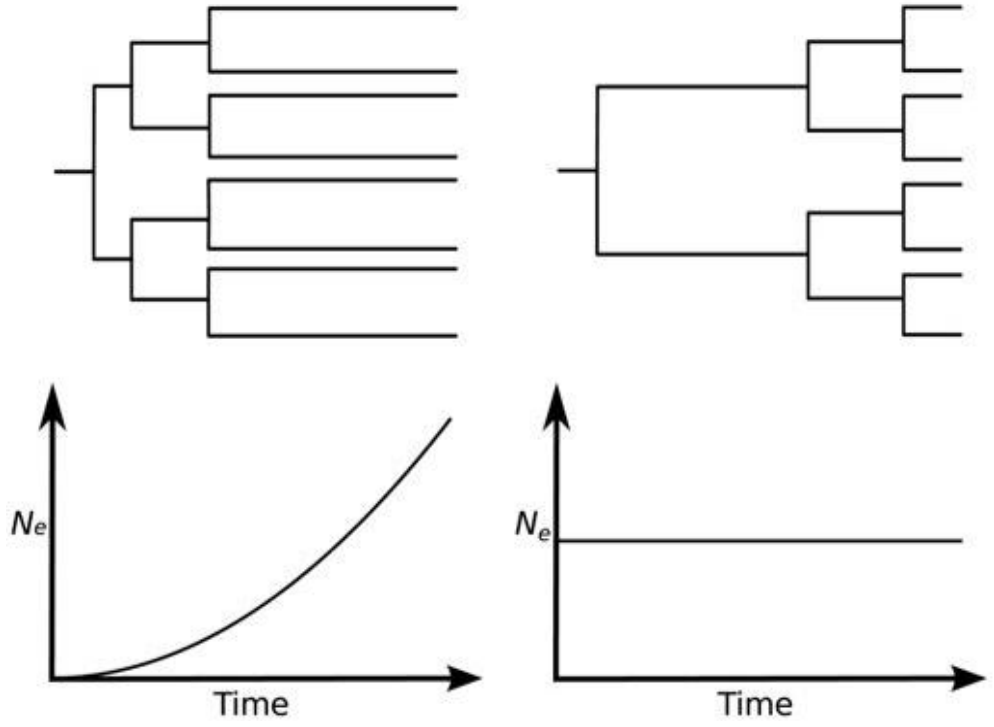
Assume that the strict clock is a relaxed clock with very low rate variation (it is nested)!







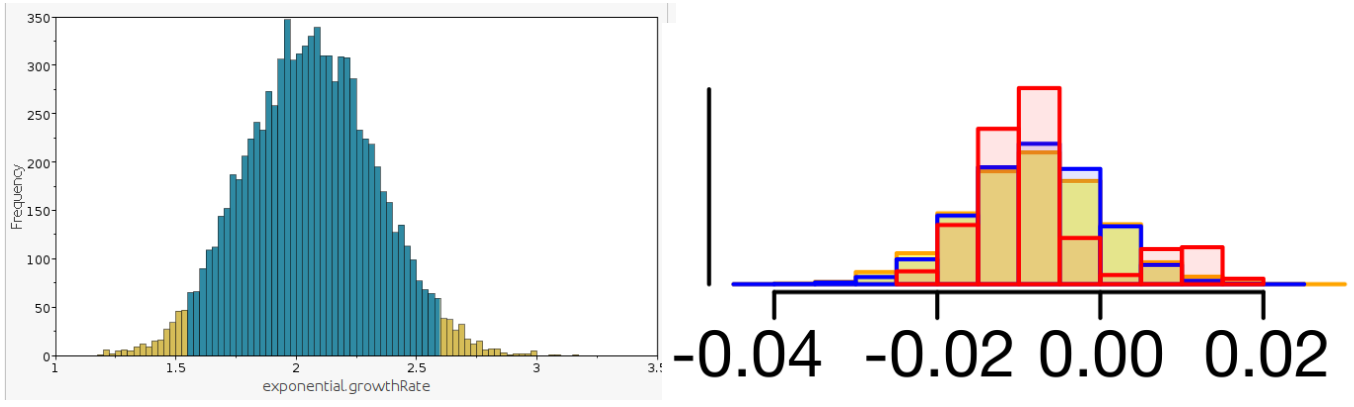
From Duchene et al. 2018

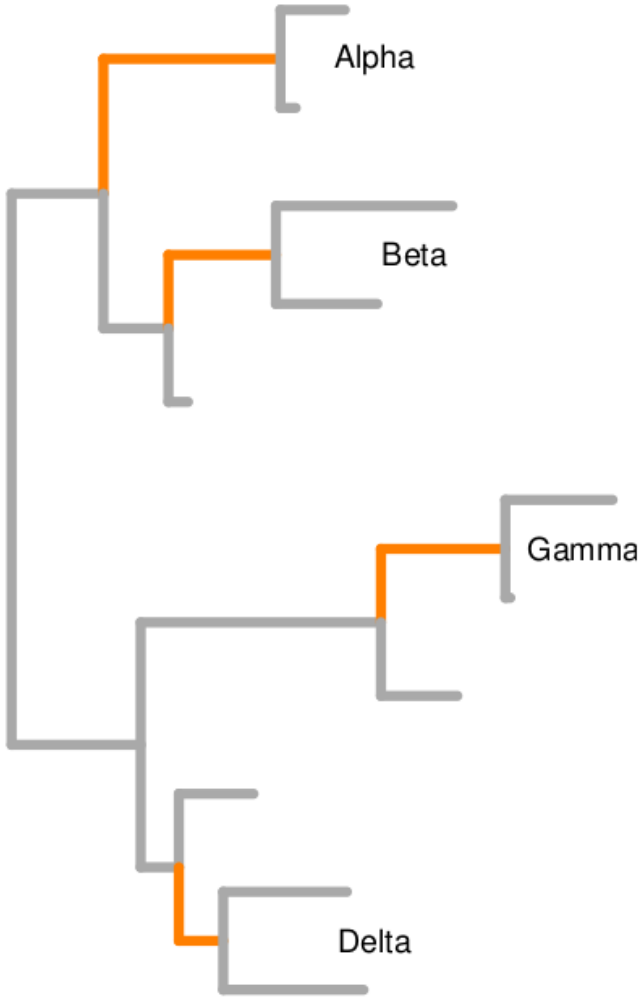


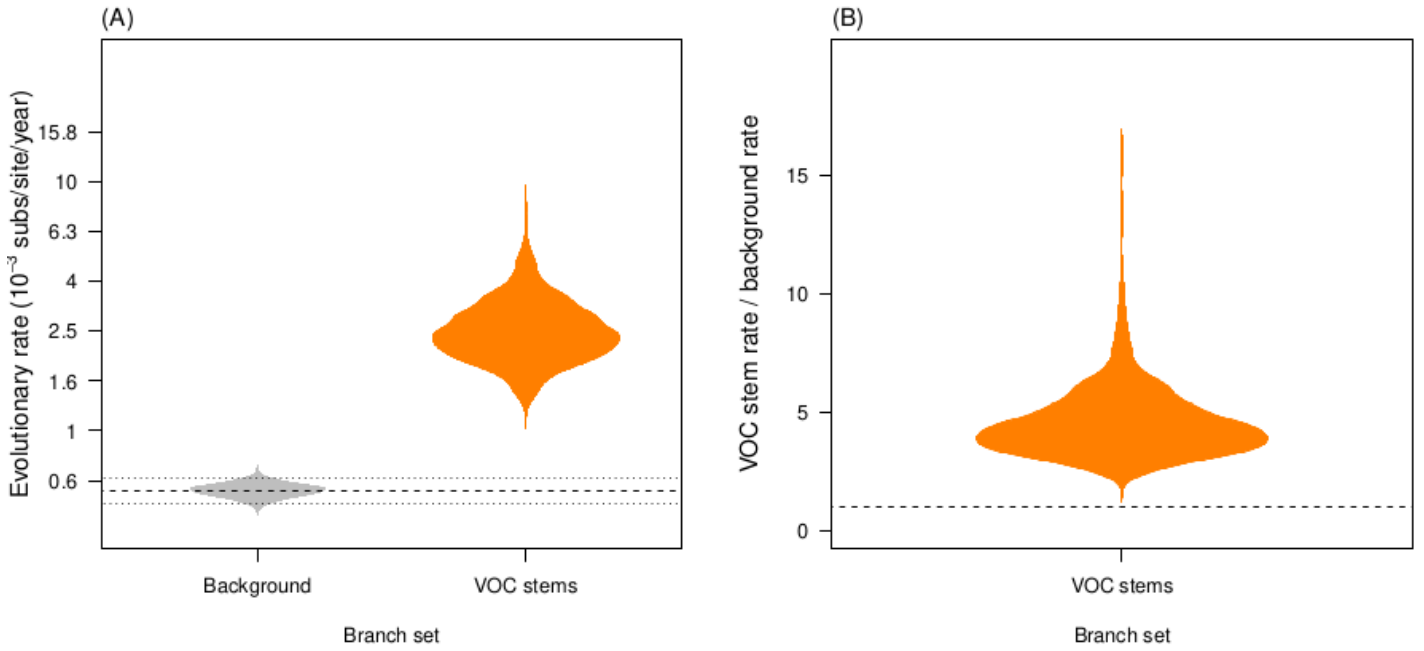
Under the exponential growth:

$N_e = N_0 * e^{rt}$
For $r > 0$

$N_e = N_0 * e^{rt}$
If $r = 0$
$N_e = N_0$





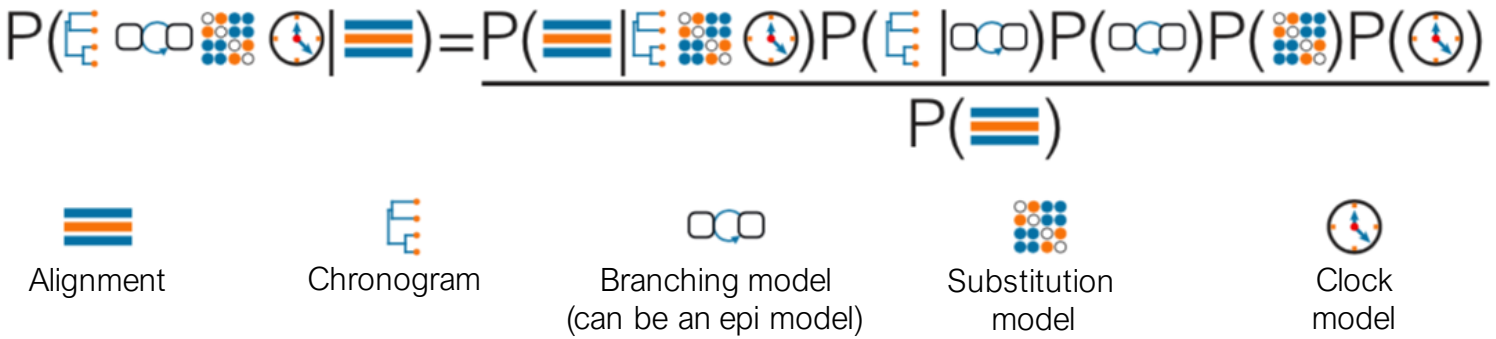


From: Tay et al. 2021

Concept summary

- For some models, parameters or statistics can give an idea of how well it is supported, **relative to nested models**.

Key traces



- HKY+G: 2 parameters (kappa, shape)
- Coalescent exponential: 2 parameters (growthRate, ePopSize)
- UCLN clock model: 2 parameters (ucldMean, ucldStdev)*

From: Du Plessis and Stadler 2015

* arguably the branch lengths are also parameters, but they are much less tractable.

Probability densities

Traces:

Statistic	Mean	ESS	Ty...
posterior	-41357.667	229	R
likelihood	-41378.718	1359	R
prior	21.051	216	R
treeLikelihood	-41378.718	1359	R
TreeHeight	0.178	352	R
kappa	4.94	1753	R
gammaShape	0.47	349	R
ucldMean	1.236E-3	301	R
ucldStdev	0.272	537	R
rate.mean	1.203E-3	285	R
rate.variance	4.245E-7	679	R
rate.coefficientOfVa...	0.29	518	R
CoalescentExponent...	32.865	215	R
ePopSize	13.162	350	R
growthRate	36.71	304	R

Summary statistics

Traces:

Statistic	Mean	ESS
posterior	-41357.667	229
likelihood	-41378.718	1359
prior	21.051	216
treeLikelihood	-41378.718	1359
TreeHeight *	0.178	352
kappa	4.94	1753
gammaShape	0.47	349
ucldMean	1.236E-3	301
ucldStdev	0.272	537
rate.mean	1.203E-3	285
rate.variance	4.245E-7	679
rate.coefficientOfVa...	0.29	518
CoalescentExponent...	32.865	215
ePopSize	13.162	350
growthRate	36.71	304

Parameters of the model *

Traces:

Statistic	Mean	ESS	Ty...
posterior	-41357.667	229	R
likelihood	-41378.718	1359	R
prior	21.051	216	R
treeLikelihood	-41378.718	1359	R
TreeHeight	0.178	352	R
kappa	4.94	1753	R
gammaShape	0.47	349	R
ucldMean	1.236E-3	301	R
ucldStdev	0.272	537	R
rate.mean	1.203E-3	285	R
rate.variance	4.245E-7	679	R
rate.coefficientOfVa...	0.29	518	R
CoalescentExponent...	32.865	215	R
ePopSize	13.162	350	R
growthRate	36.71	304	R

Alignment

Chronogram

Branching model
(can be an epi model)

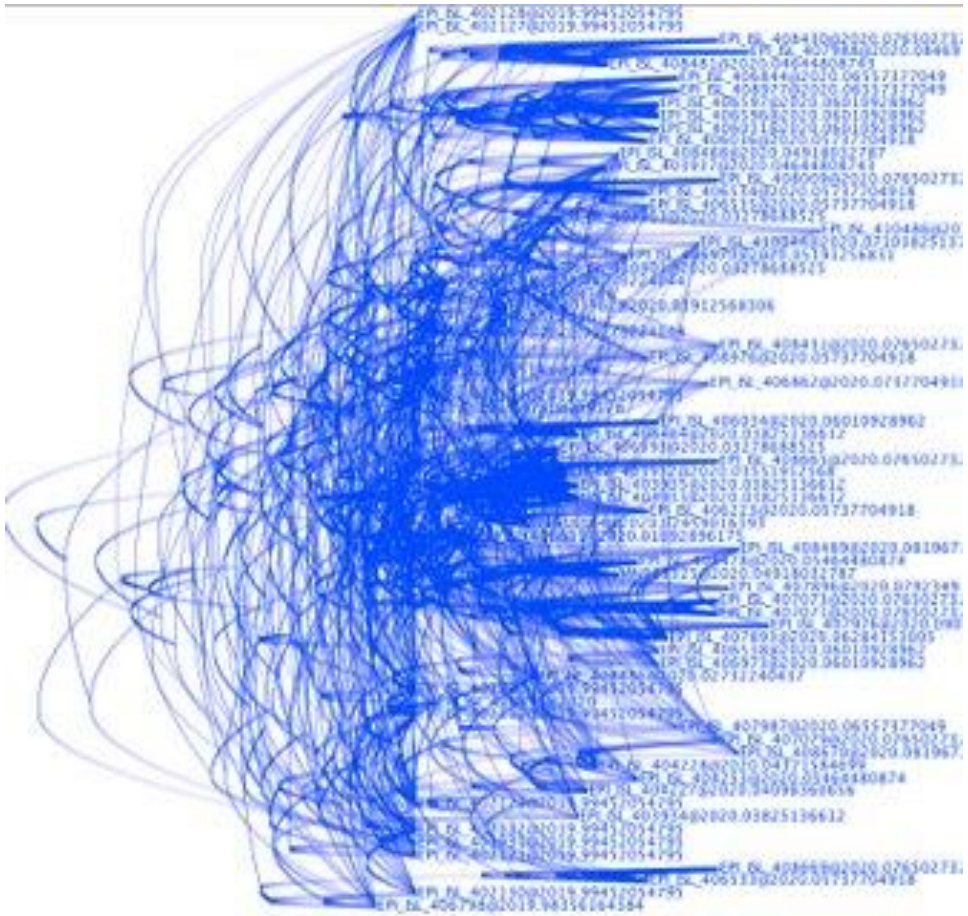
Substitution
model

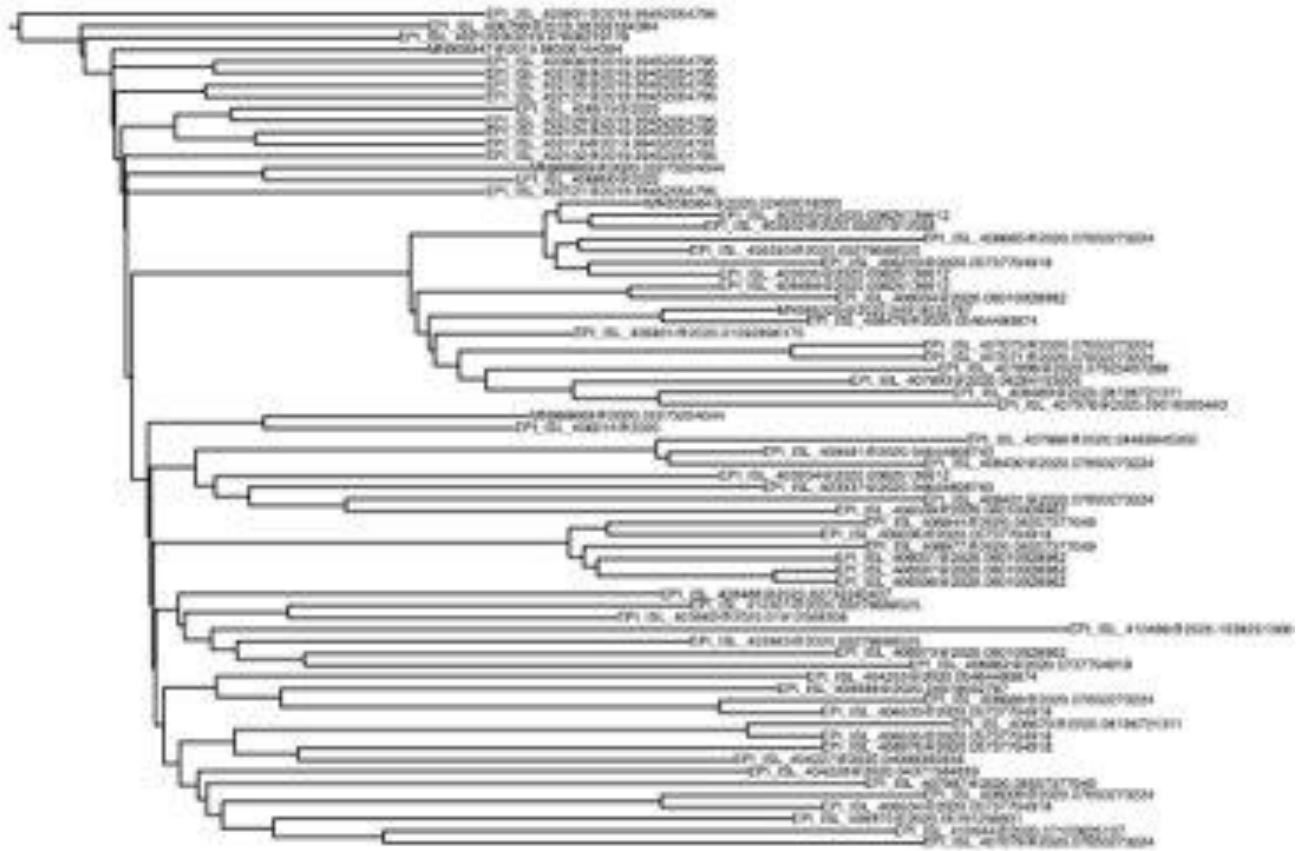
Clock
model

Concept summary

- Model parameters can be found in the posterior function, P(M, par|data).
- Arguably, the tree is also a parameter, but we rarely inpsect its trace (use AWTY).
- Summary statistics can be calculated from the model parameters.
- Probability densities are in log units and should not be compared between data sets, or used to select models.
  - The 'posterior' trace is an unnormalised posterior and NOT the posterior probability.

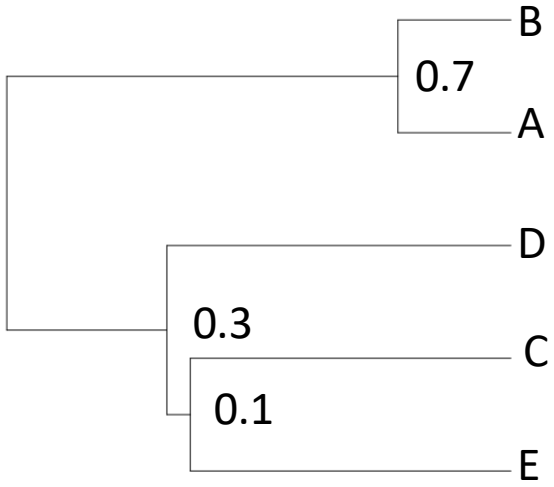
Summarising trees



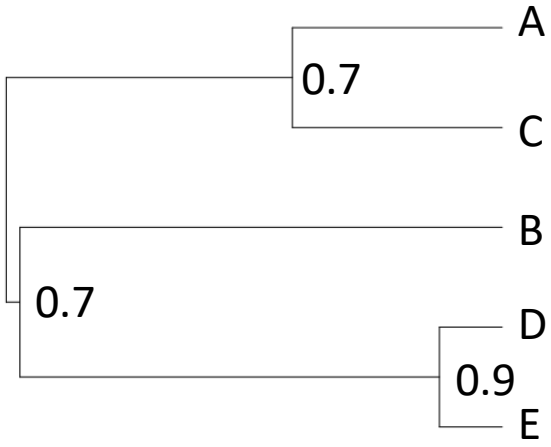


The highest clade credibility tree

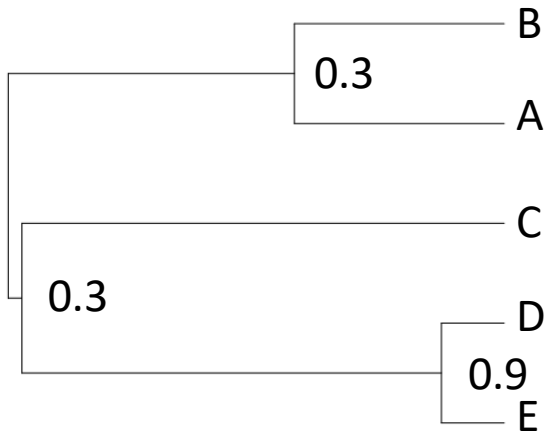
- The highest product of node posterior probabilities



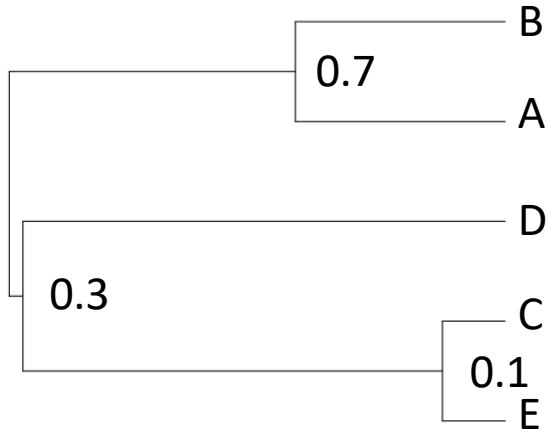
1 tree



7,000 trees



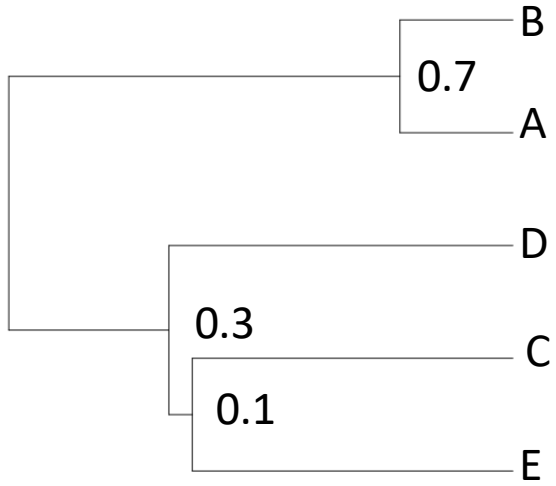
2,000 trees



999 trees

The highest clade credibility tree

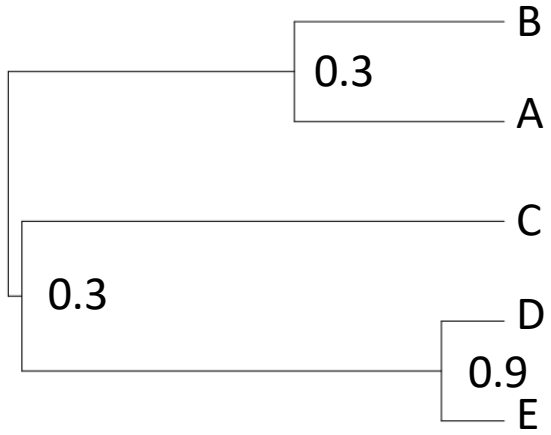
- The highest product of node posterior probabilities



1 tree 0.7*0.3*0.1 = 0.021

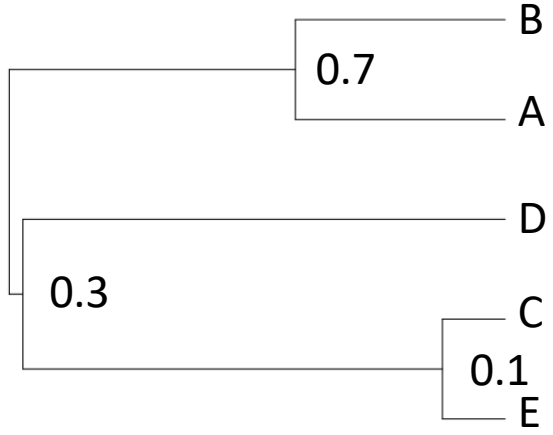
7,000 trees

0.7*0.7*0.9 = 0.44



2,000 trees

0.3*0.3*0.9 = 0.081

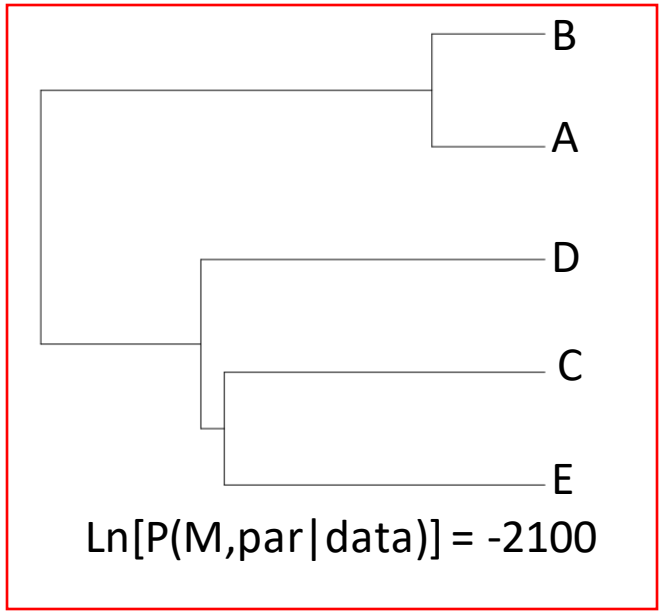


999 trees

0.7*0.3*0.1 = 0.021

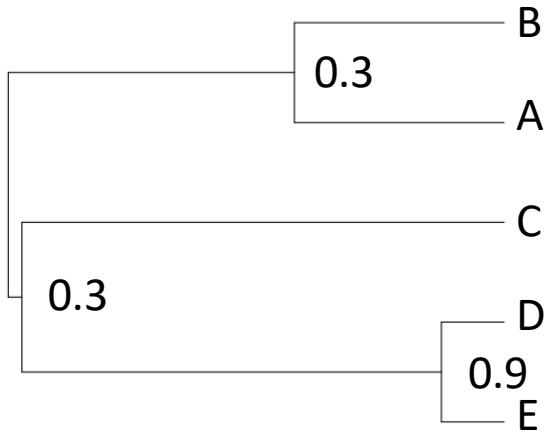
The maximum *a posteriori* tree

- The tree with highest posterior density.

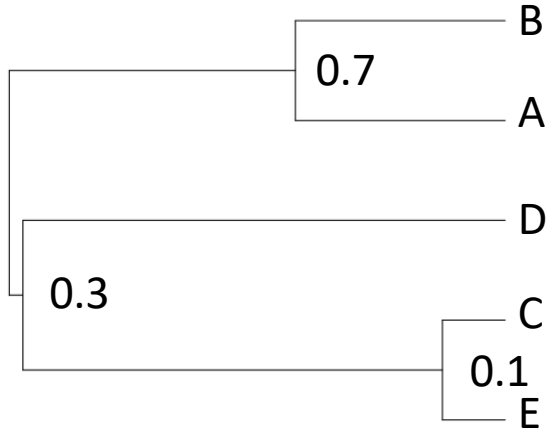


Ln[P(M,par|data)] = -2100

Ln[P(M,par|data)] = -3000



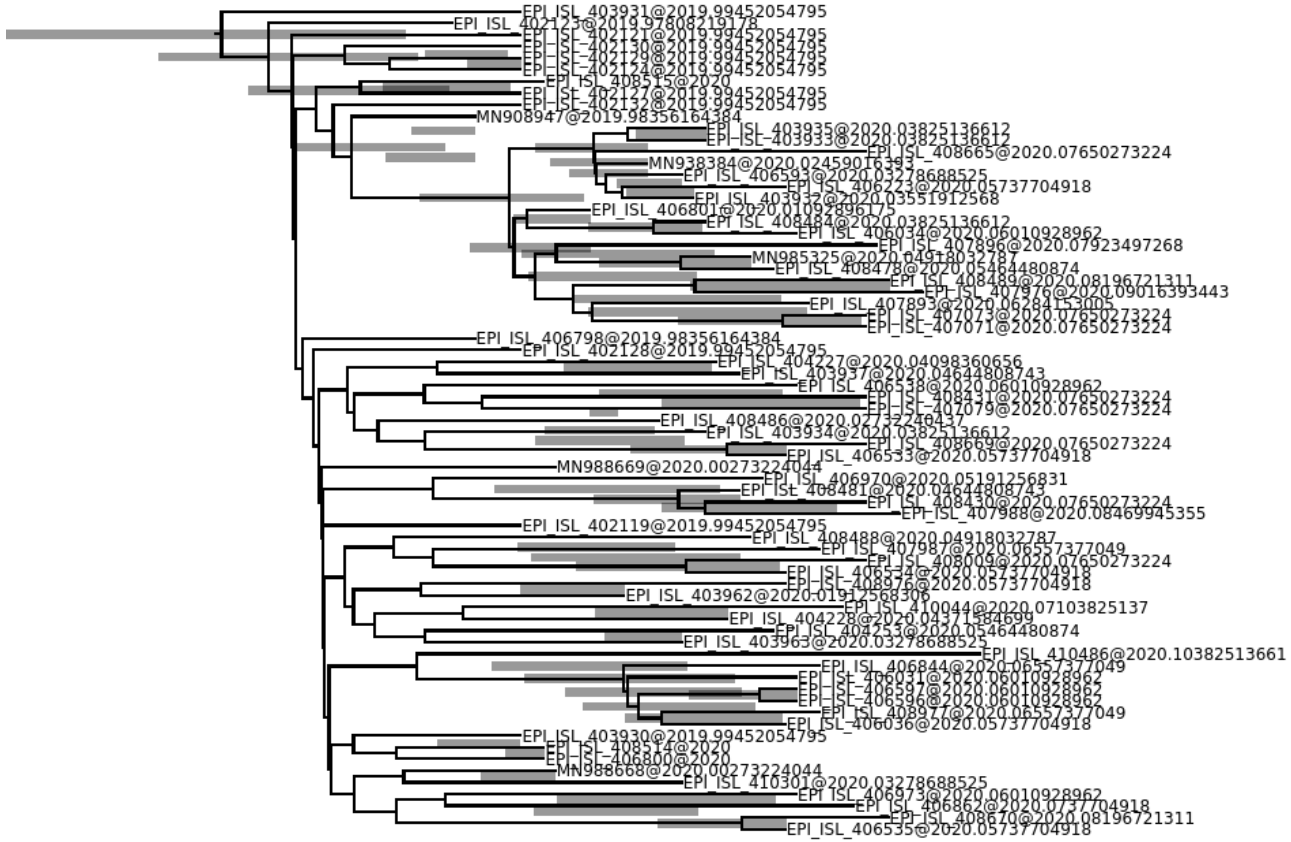
Ln[P(M,par|data)] = -4000



Ln[P(M,par|data)] = -4100

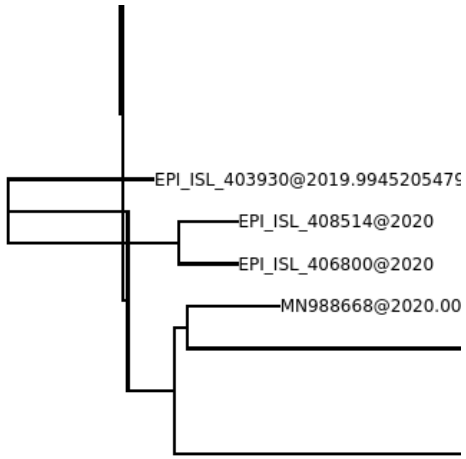
Node heights

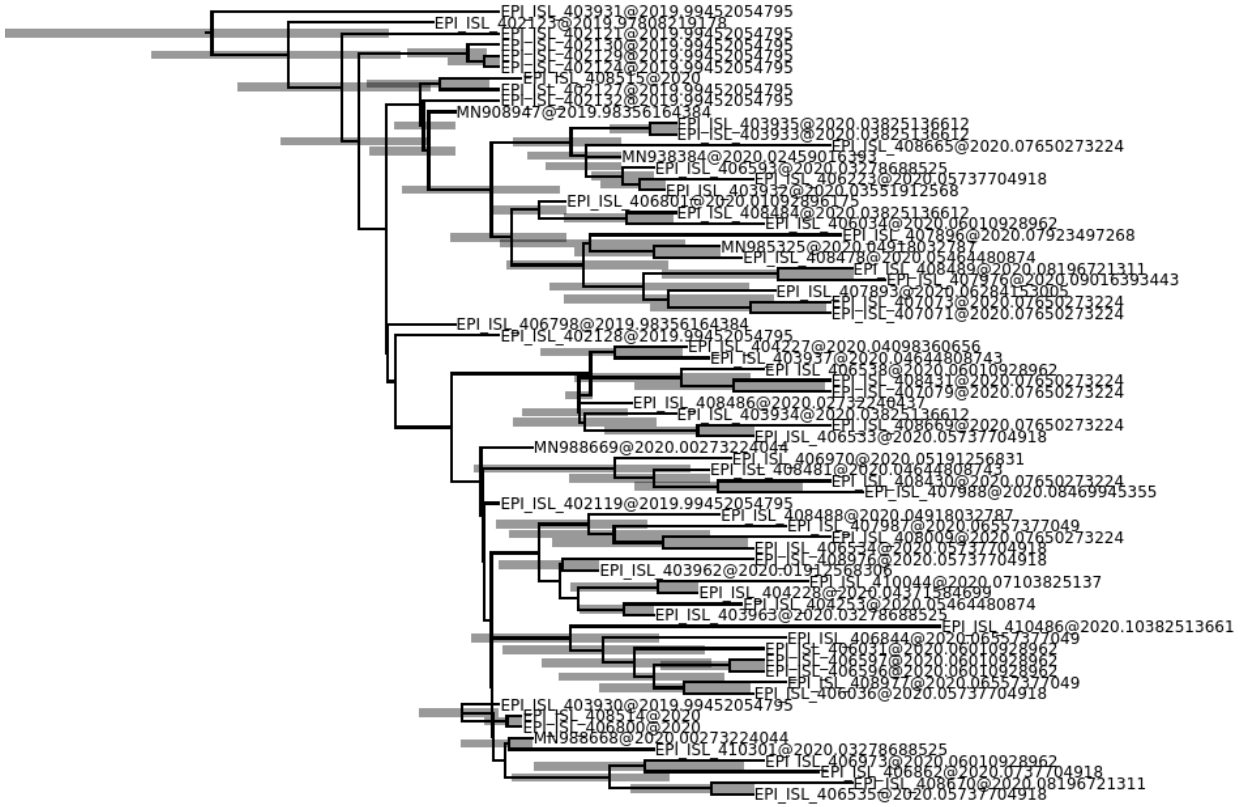
- Keep: keep node heights of summary tree.
  - Can have credible intervals outside of nodes.



Node heights

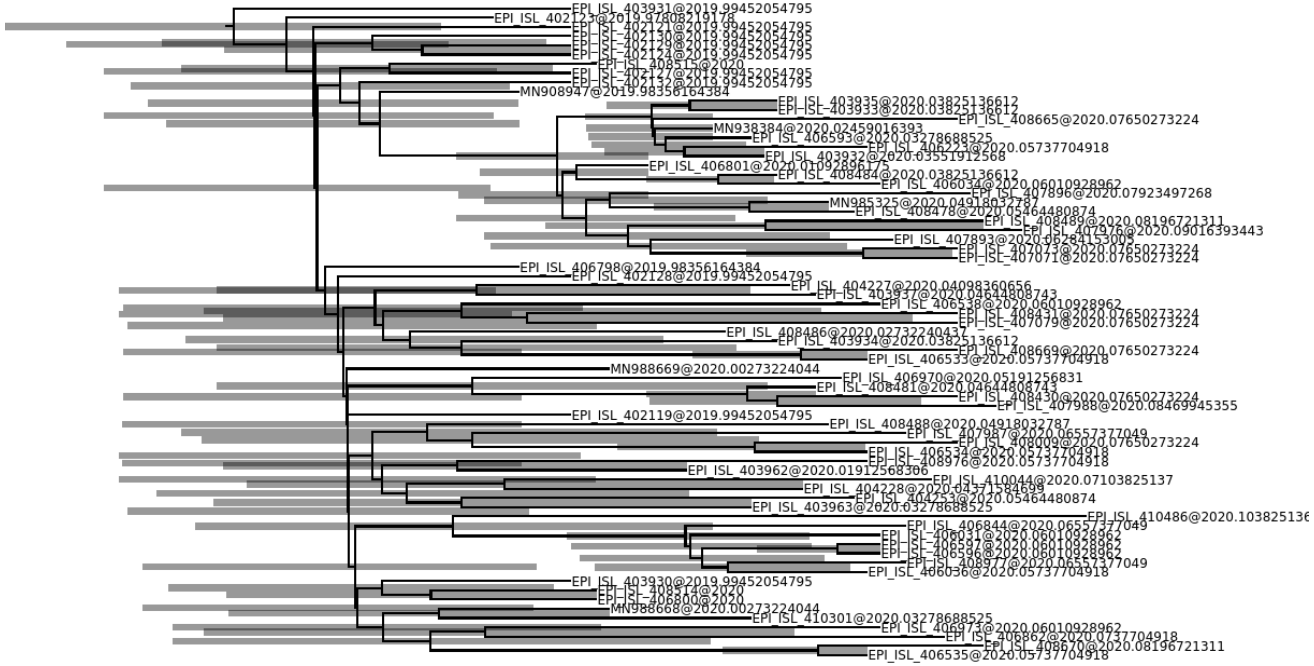
- Mean: take average of nodes if present in other trees.
  - Can have negative branch lengths, especially for poorly supported nodes.





Node heights

- Common ancestor: take the mean age of taxa for all trees (regardless of monophyly).
  - Can lead to very deep divergences and wide credible intervals.
  - No negative branch lengths.



Concept summary

- Trees can be summarised in different ways and they all have compromises.
- Approaches to summarising node times can distort branch lengths (e.g. mean or common ancestor).
- The 'best' tree is not necessarily a reasonable summary (e.g. a very wide posterior distribution).