Tip-dating and phylodynamics

Sebastian Duchene

THE UNIVERSITY OF
MELBOURNE

In this talk…

1. Measurably evolving populations
2. Assessing temporal structure
3. Bayesian methods to assess temporal structure
    1. Case studies in influenza viruses and *Mycobacterium tuberculosis*
4. A case study using ancient DNA
5. Tree priors and tip-dating
    1. Case studies in influenza and human respiratory syncytial virus
6. Model adequacy in phylodynamics
    1. Assessing the sampling process

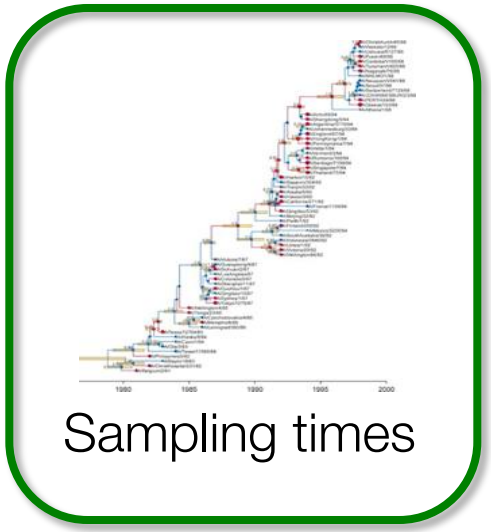
Measurably Evolving Populations

Several slides have been modified from presentations by Simon Ho

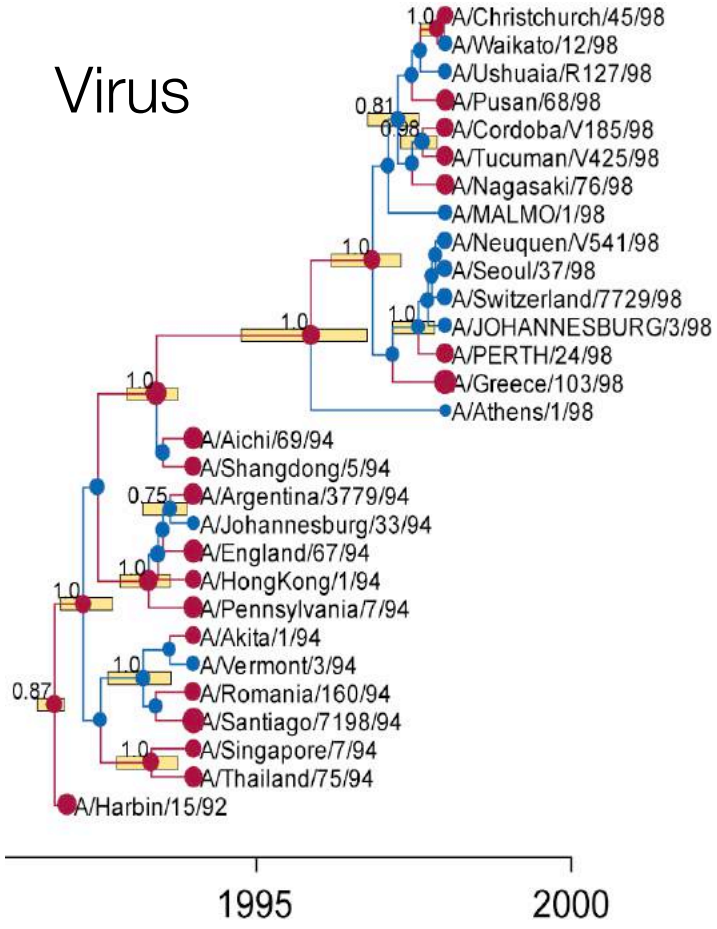
Measurably evolving populations

- Can perform tip-dating on measurably evolving populations
  - Substantial genetic change during the sampling window
  - High rate (pathogens) or wide window (ancient DNA)
- Sampling window represents large fraction of the tree height

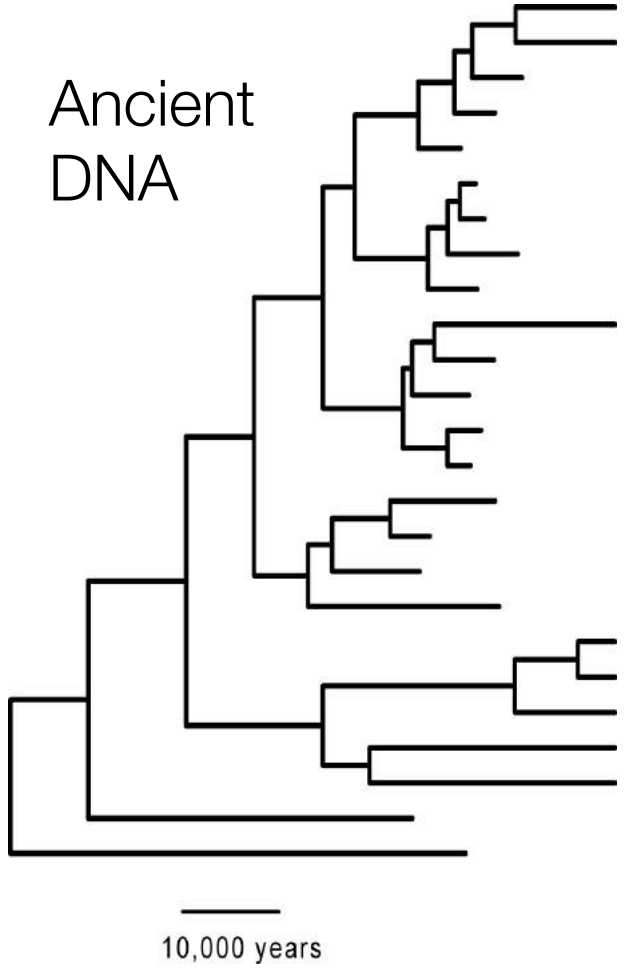
Calibration: Sampling times



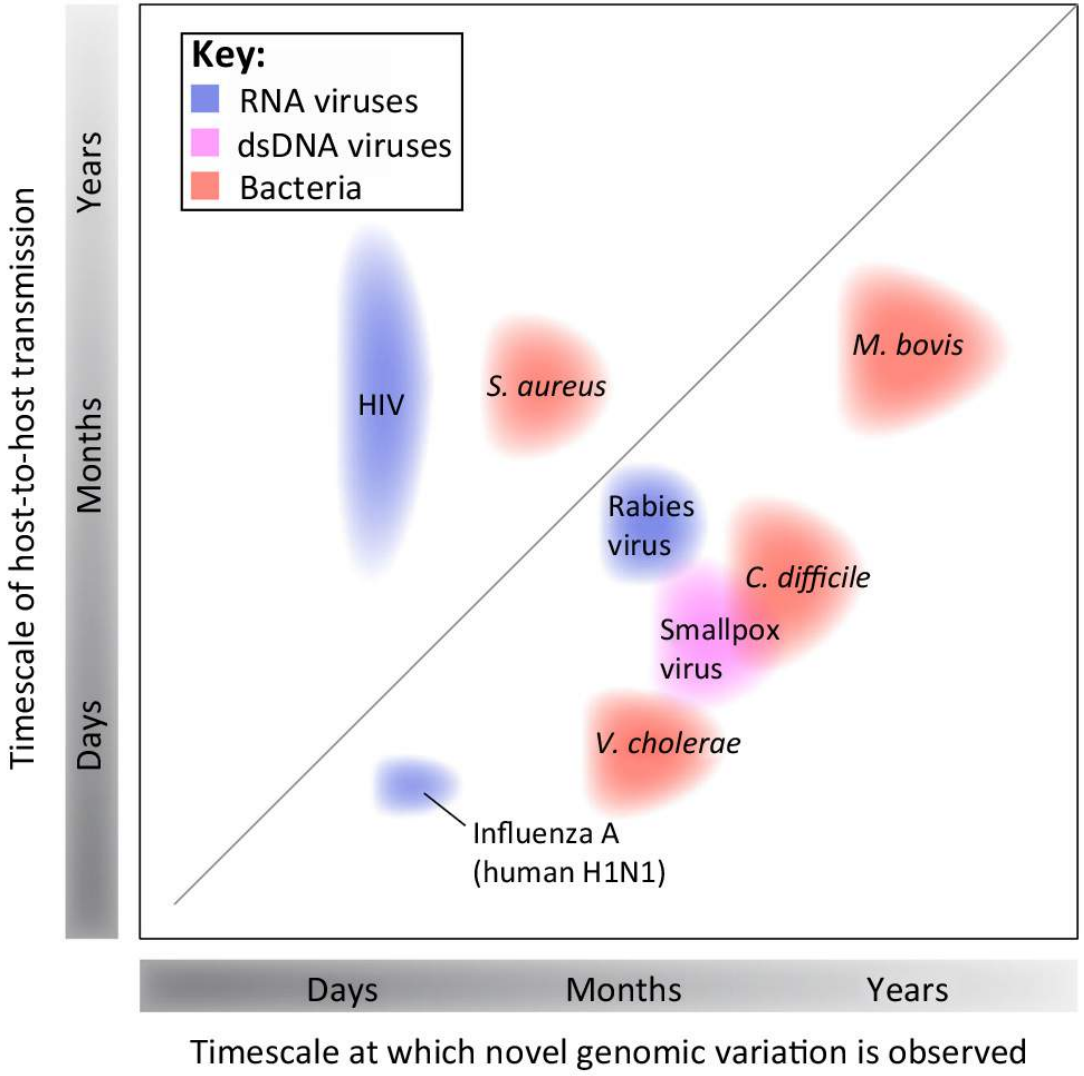
Virus



Ancient DNA

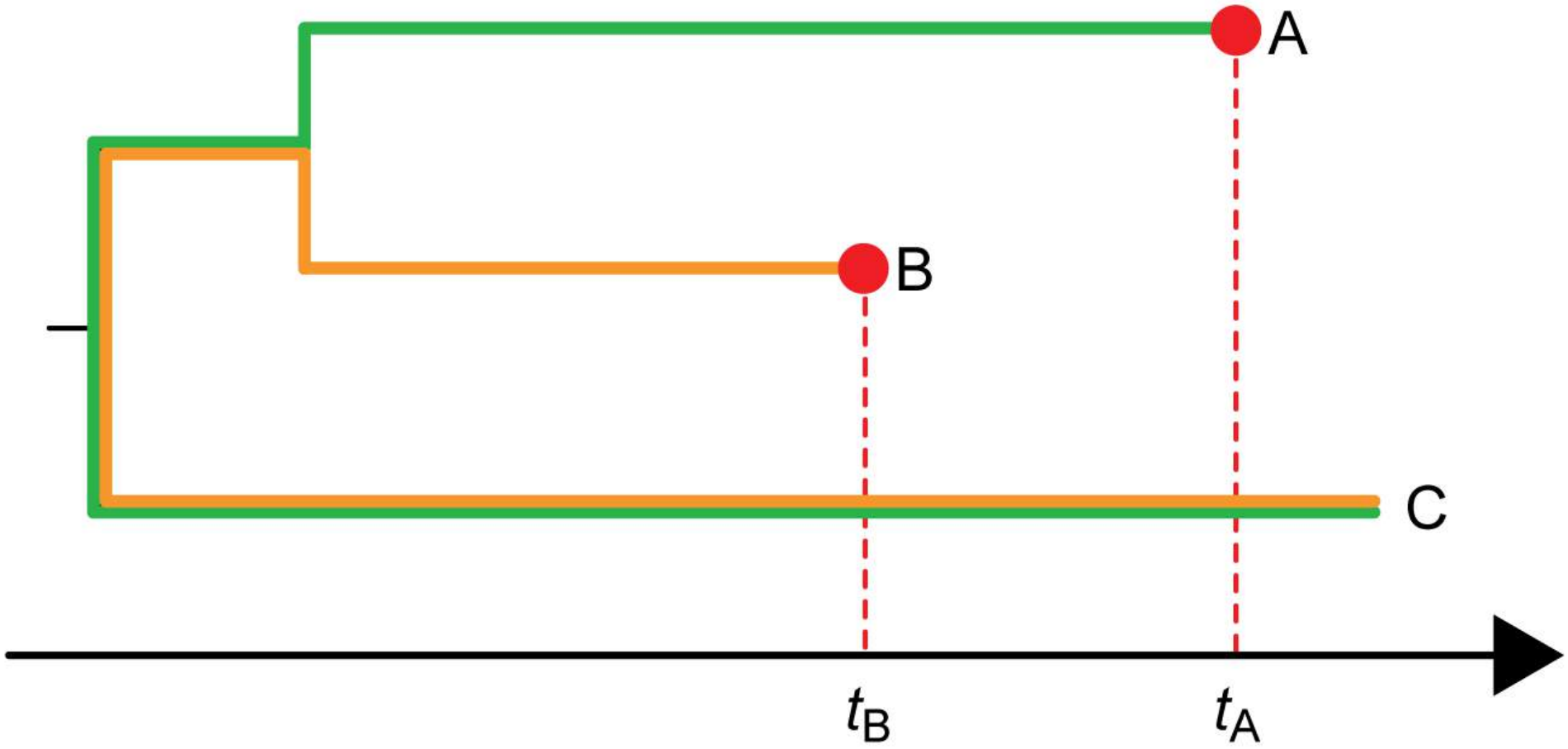


Rapidly evolving pathogens



Biek et al. (2015) Trends Ecol Evol

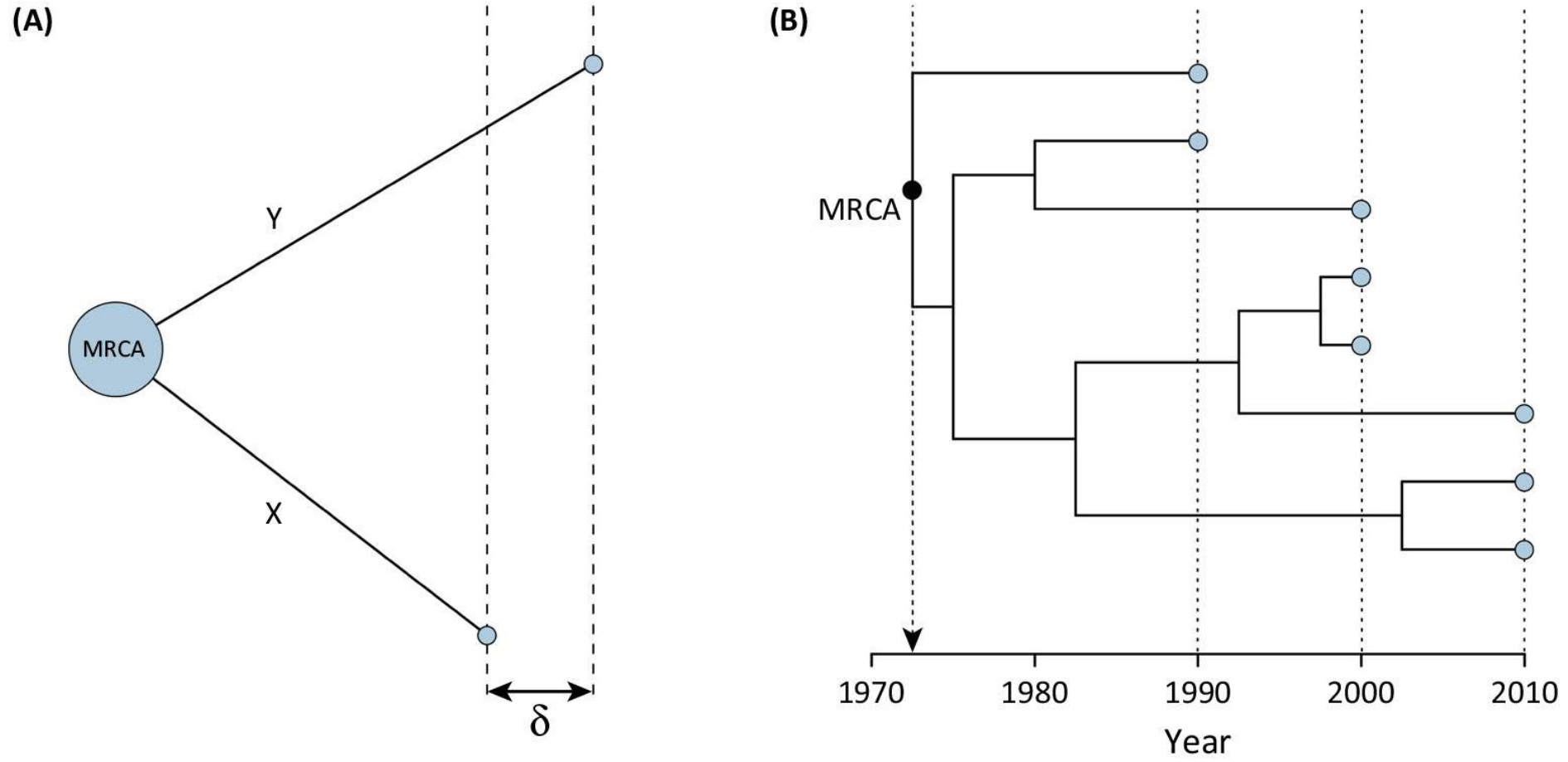
Estimating rates



$$\frac{|dB - dA|}{|tA - tB|} = \text{Substitution rate}$$

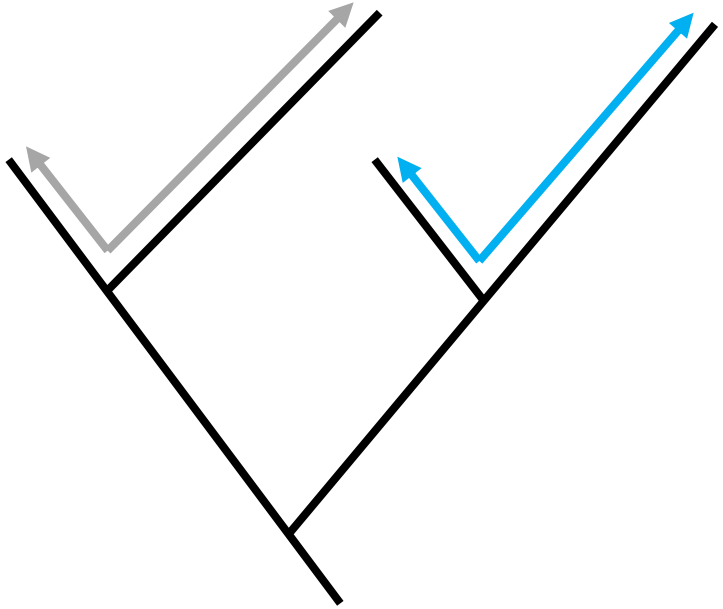
Rambaut (2000) Bioinformatics

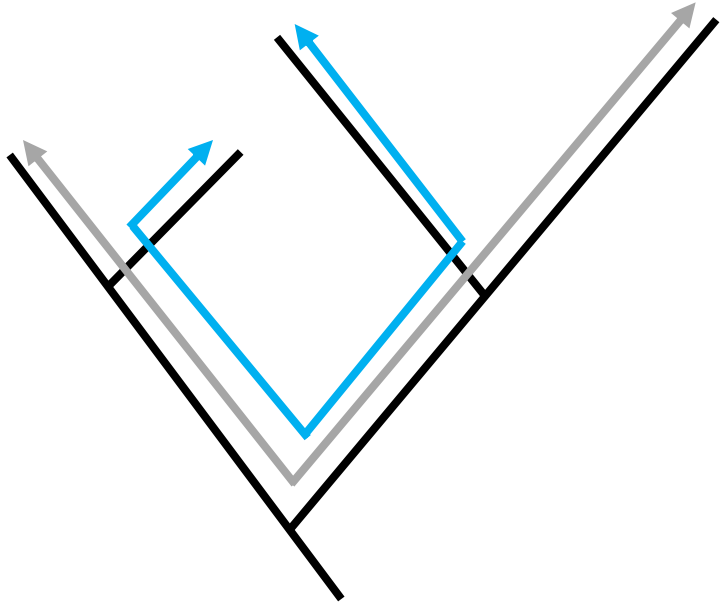
Estimating rates



Biek et al. (2015) Trends Ecol Evol

Estimating rates: phylo-temporal clustering

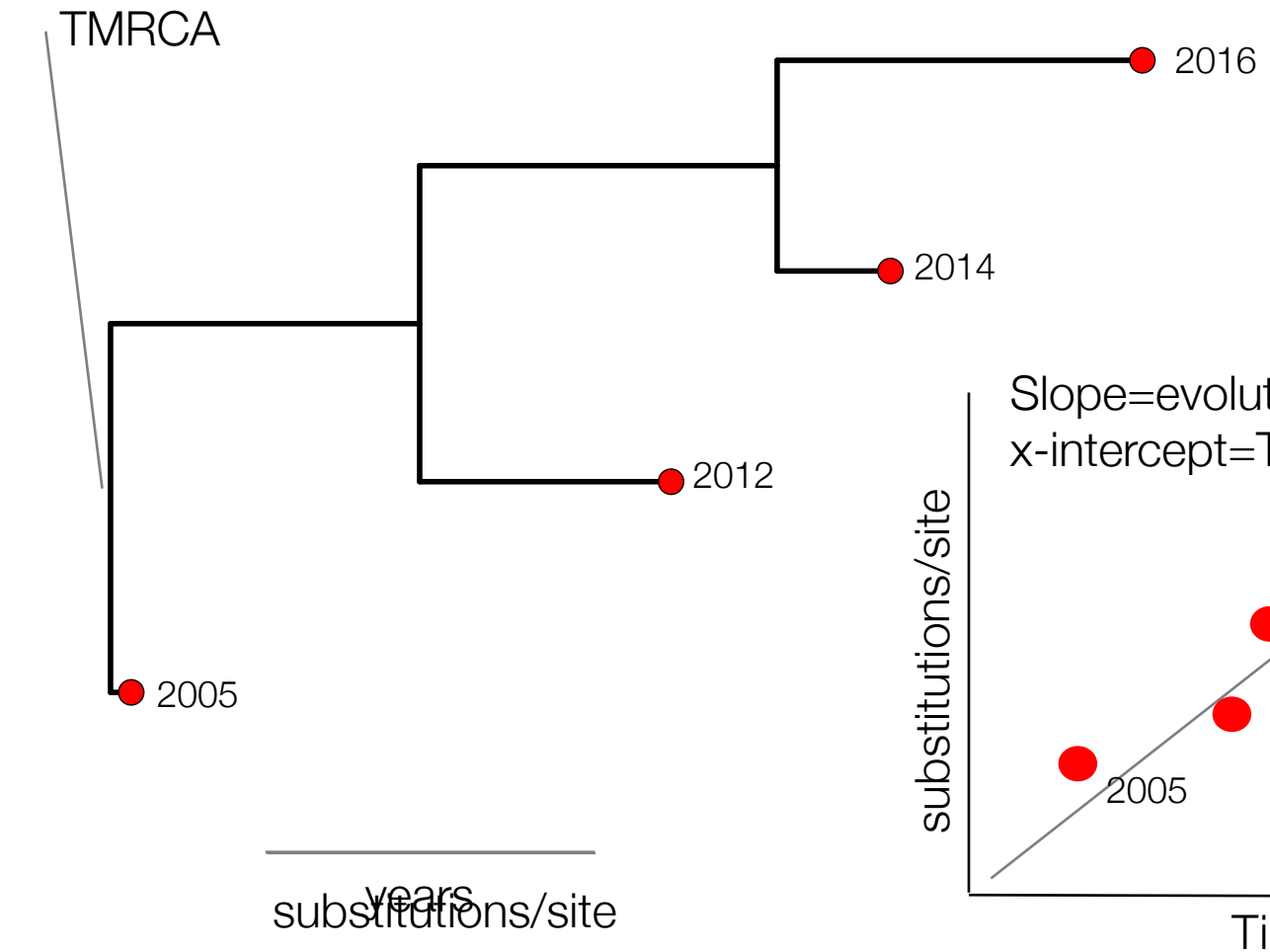




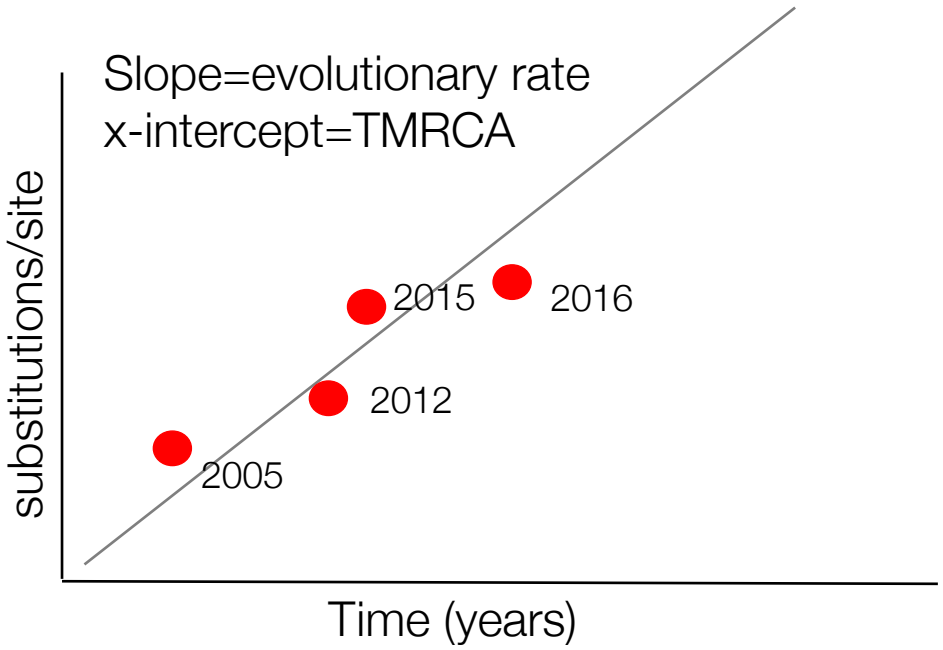
Duchêne *et al.* 2015 *Molecular Biol Evol*

Murray *et al.* 2016 *Methods in Ecol. Evol.*

Assessing temporal structure

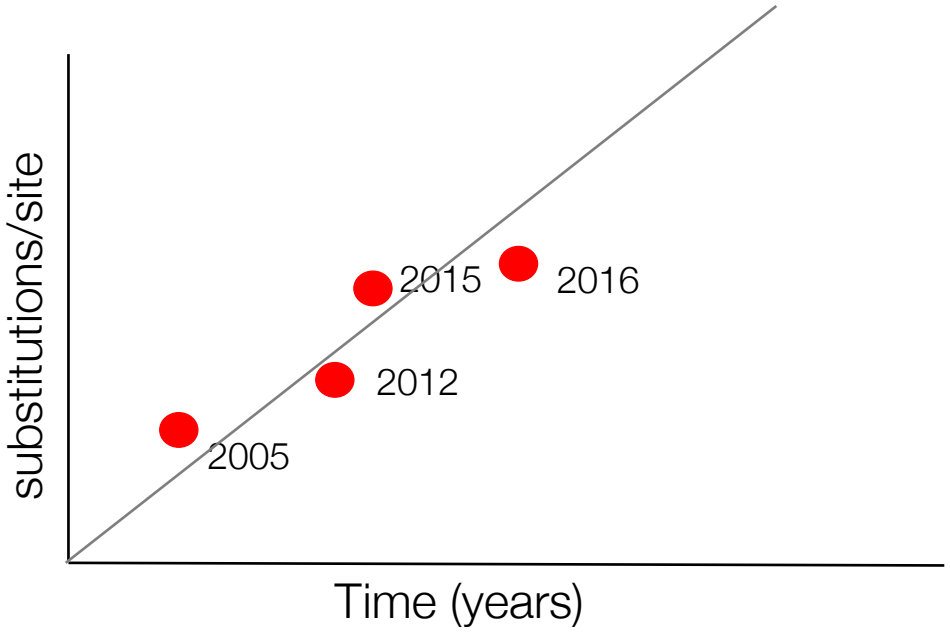


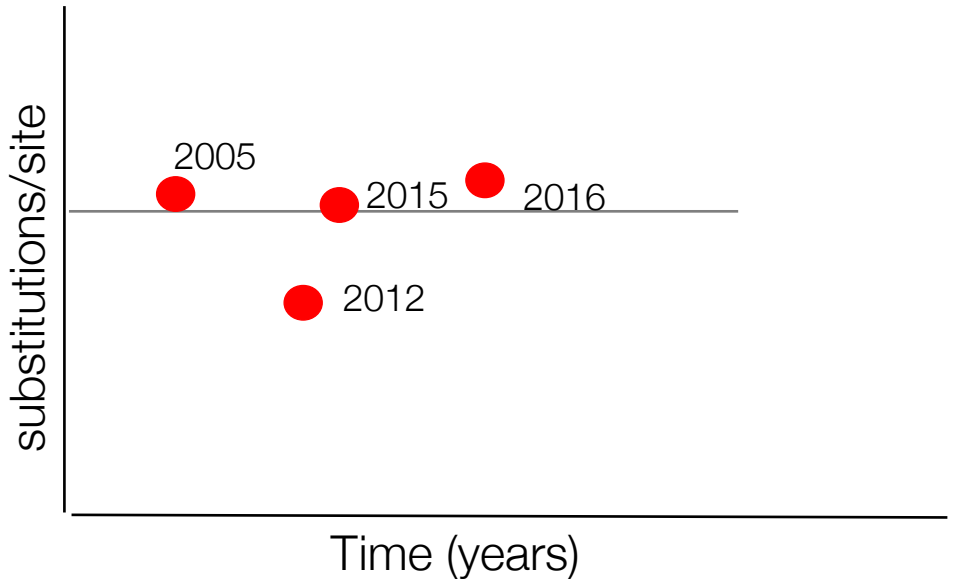
These estimates are phylogenetically non-independent!



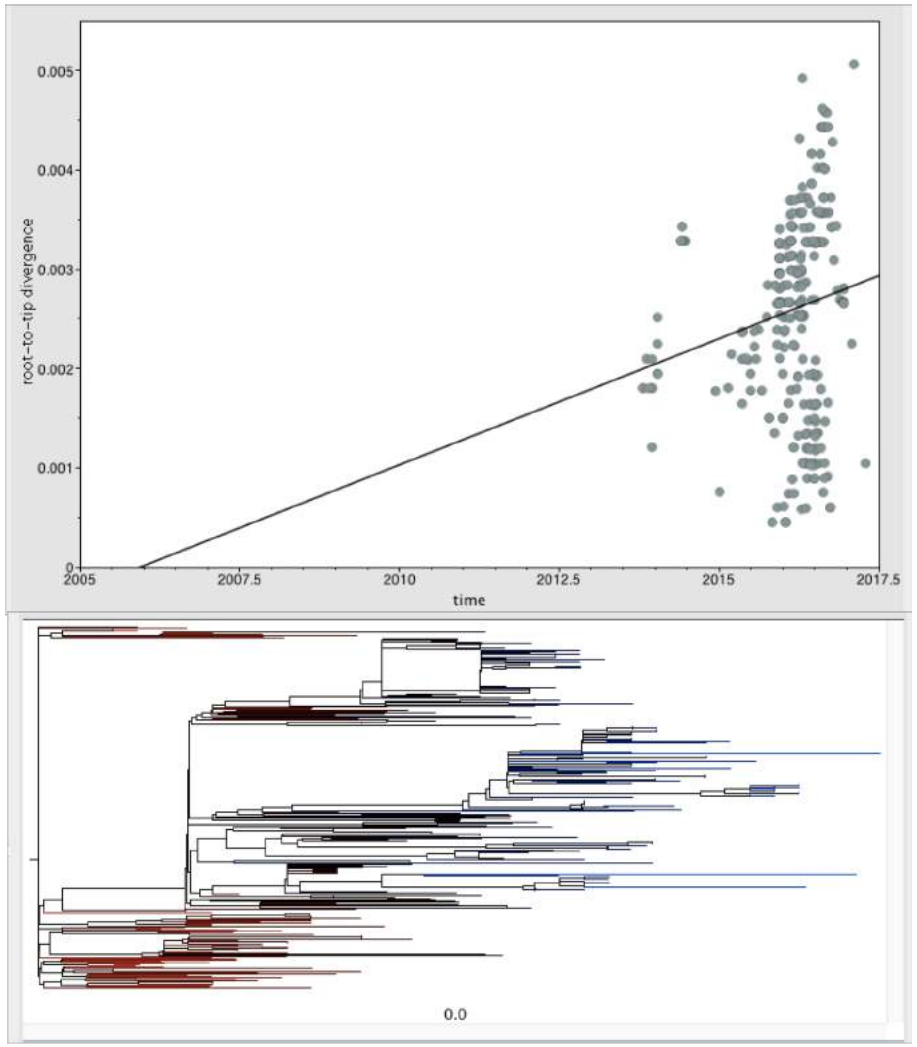
Duchêne *et al.* (2015) *Mol Biol Evol*

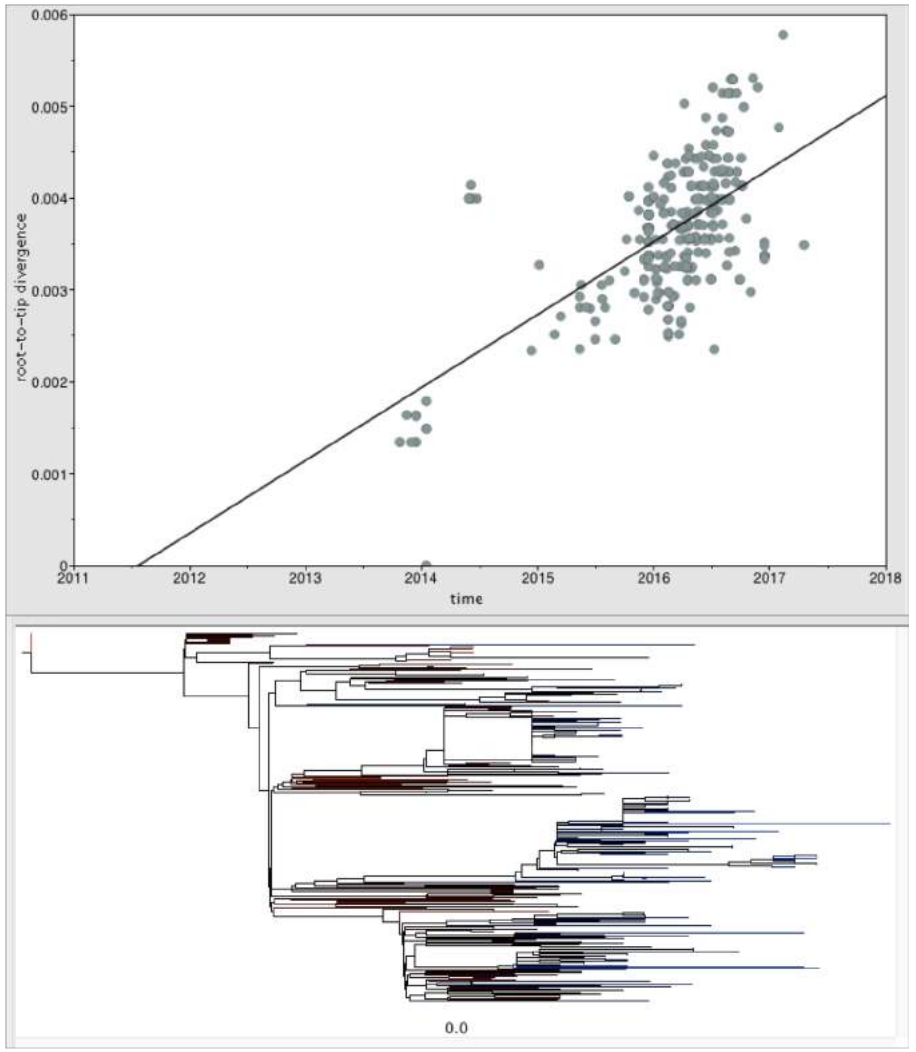
Ho and Duchêne (2014) *Mol Ecol*



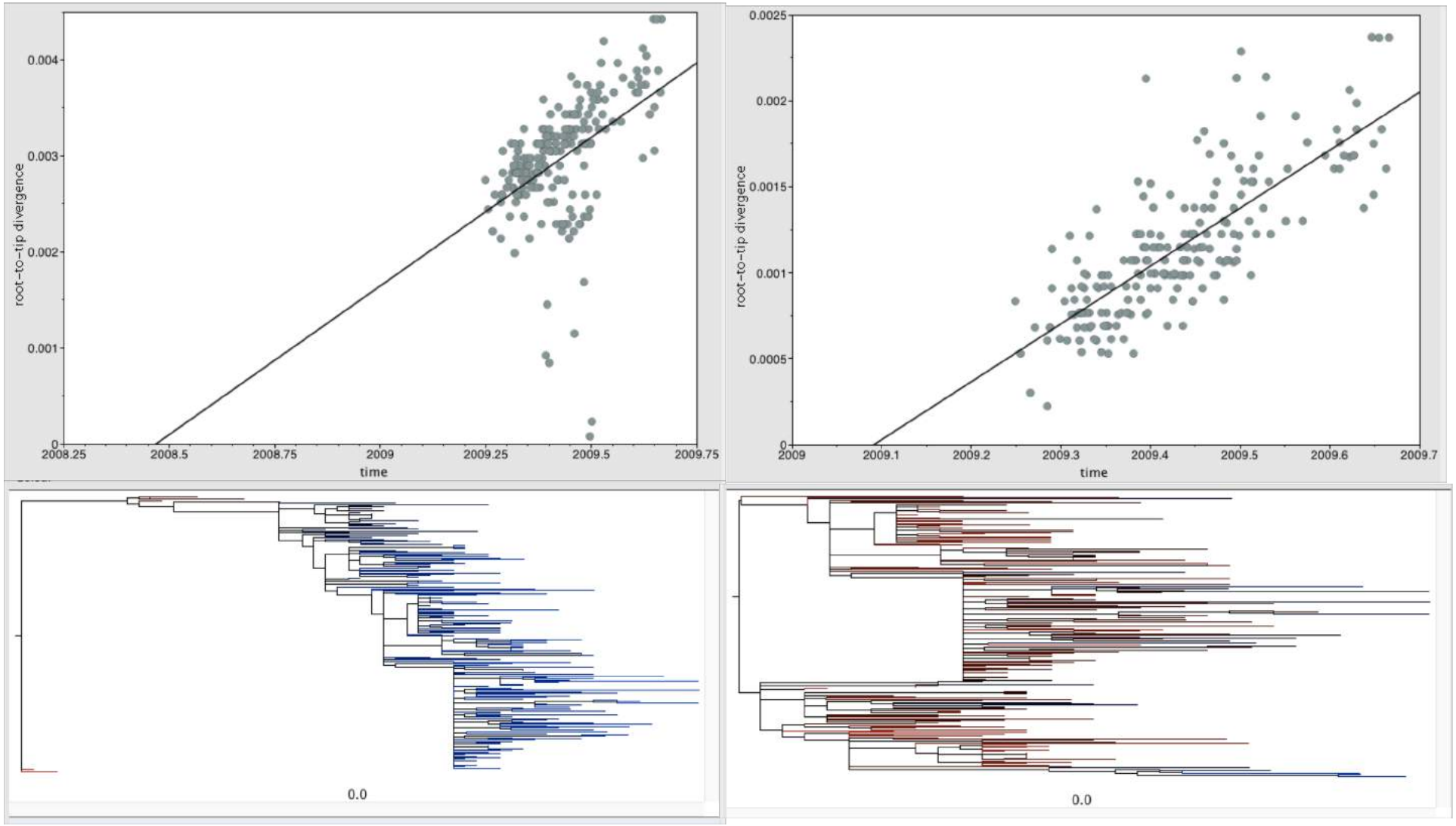


Slope should be positive
Hight R^2 (i.e. clocklike behavior)
Do not use p-value!

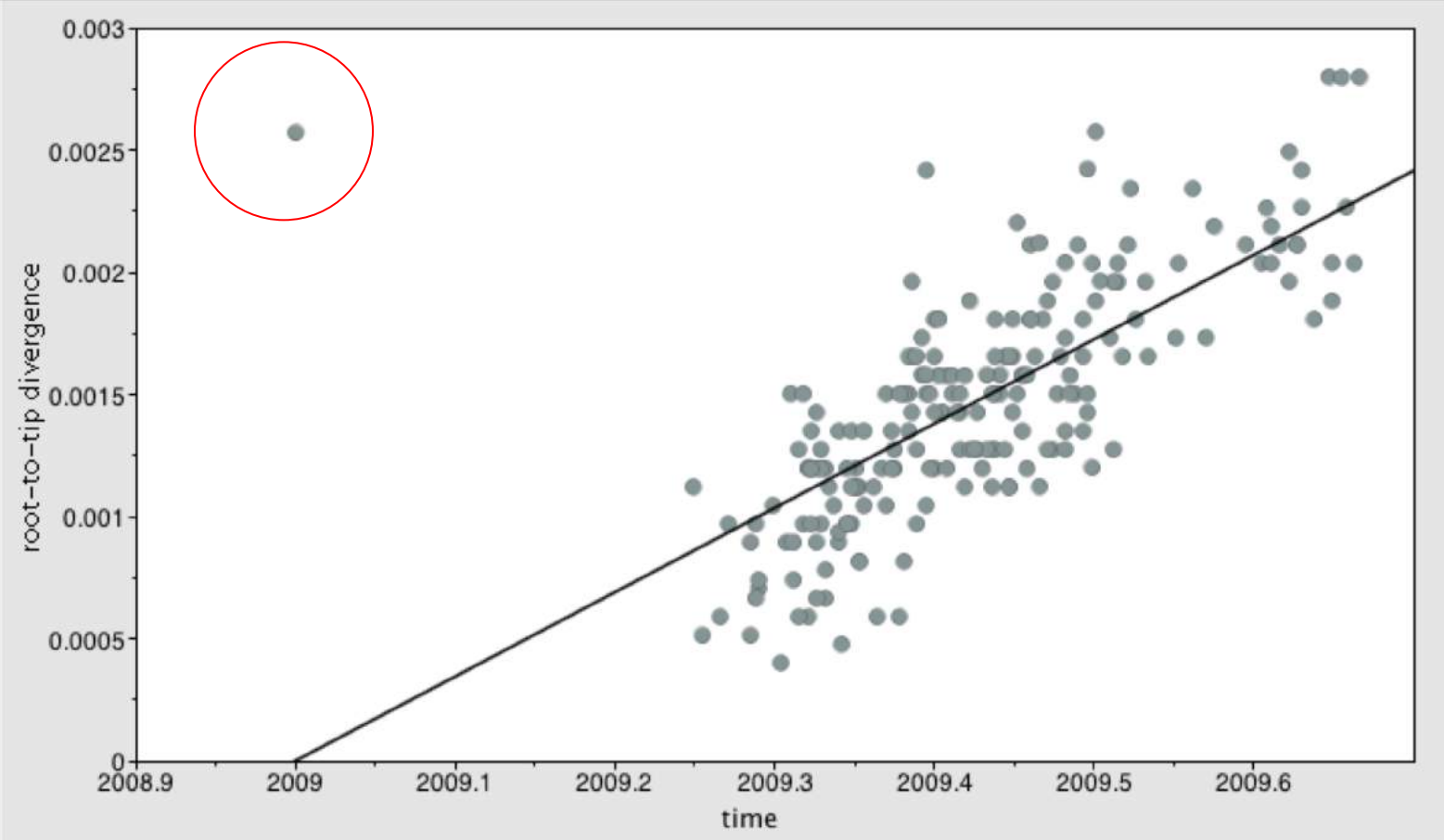




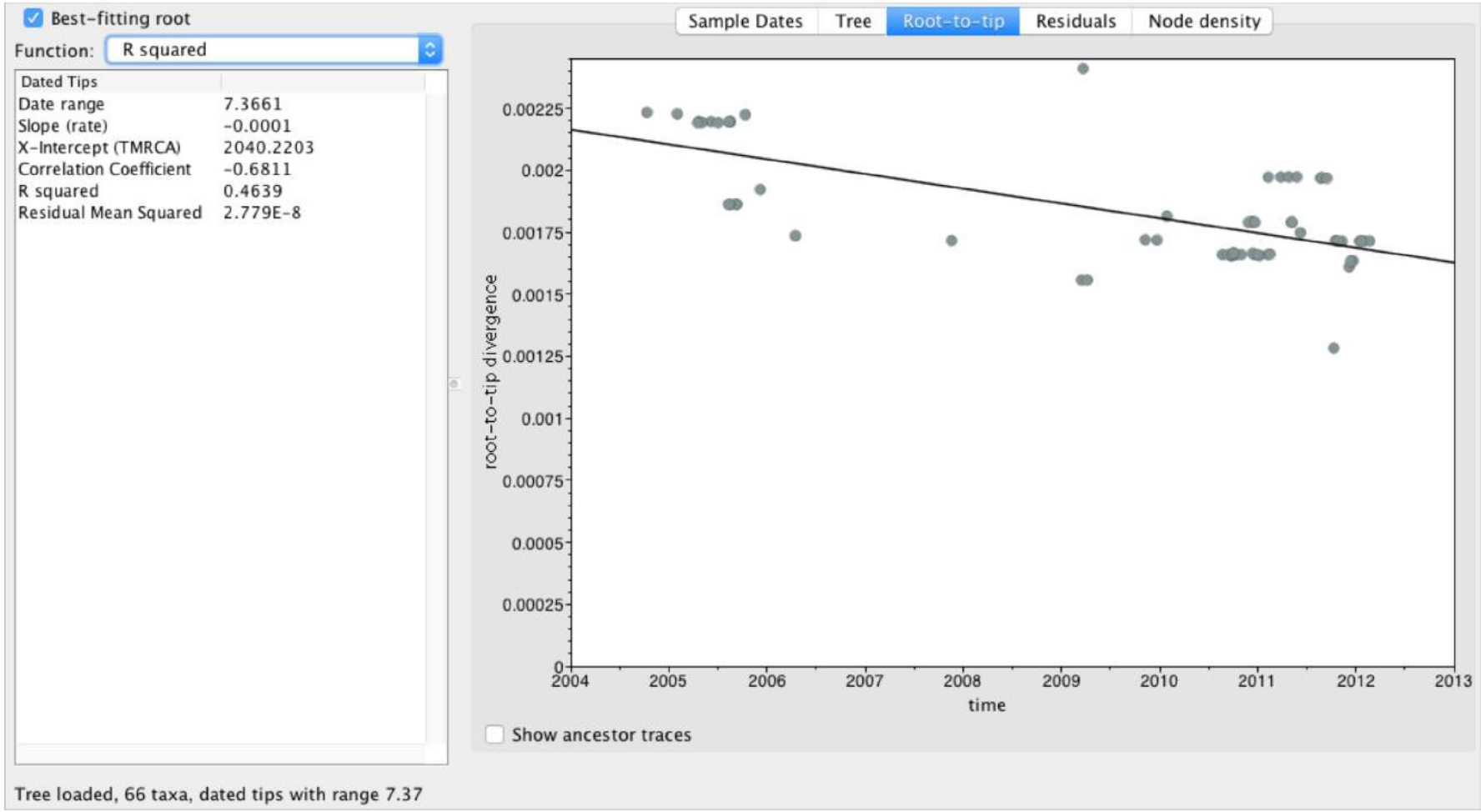
Zika virus in the Americas


root-to-tip divergence
time
0.004
0.003
0.002
0.001
0
2008.25
2008.5
2008.75
2009
2009.25
2009.5
2009.75
0.0025
0.0015
0.0005
2009.1
2009.2
2009.3
2009.4
2009.6
2009.7
0.0

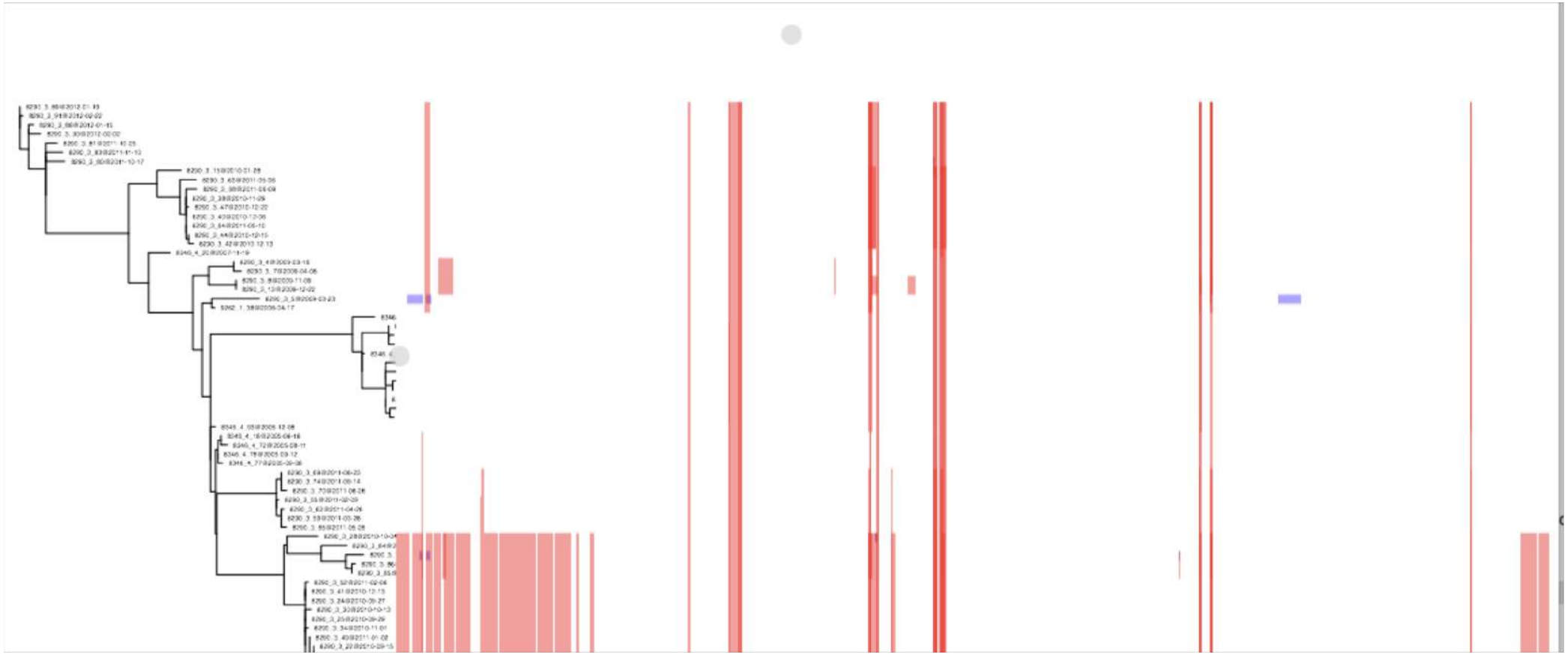
H1N1 flu

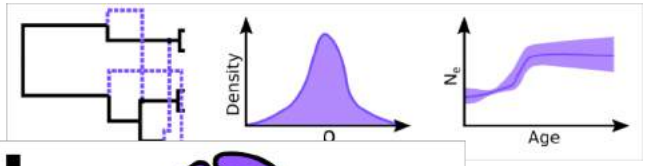

0.003
0.0025
0.002
0.0015
0.001
0.0005
0
root-to-tip divergence
2008.9
2009
2009.1
2009.2
2009.3
2009.4
2009.5
2009.6
time

Outliers?



Acinetobacter baumanii





bacter

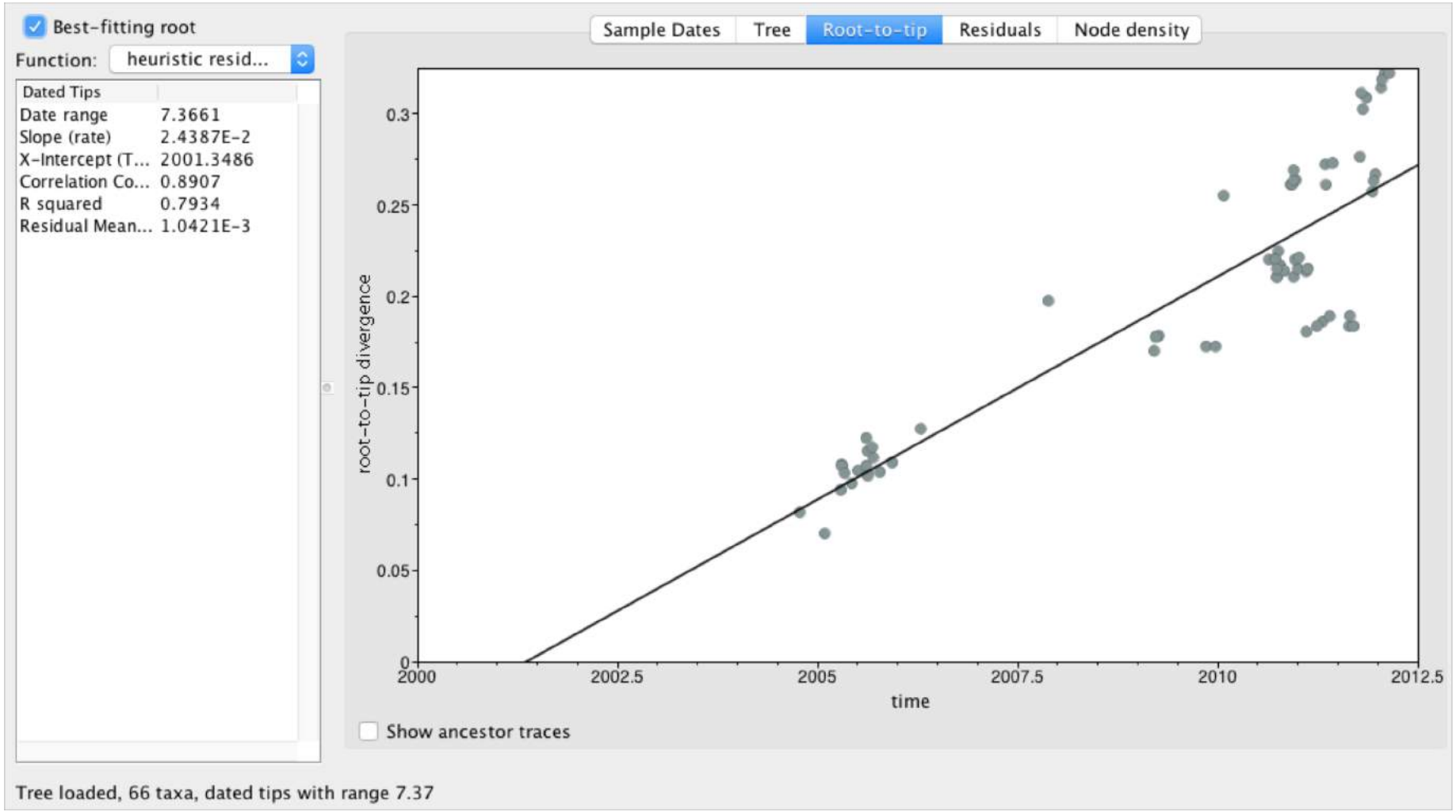
Vaughan et al. 2017

Gubbins

Genealogies Unbiased By recomBinations In Nucleotide Sequences

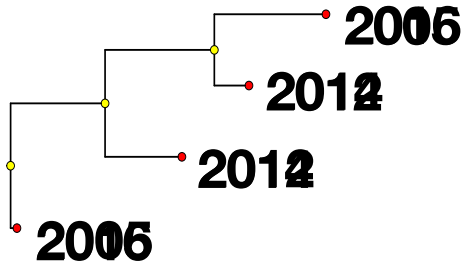
Croucher et al. 2014

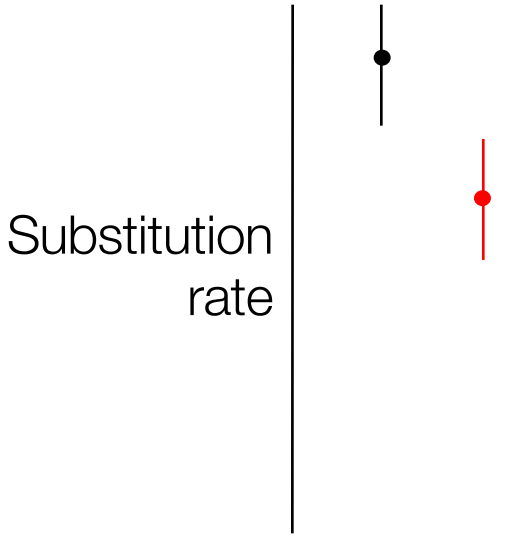
Acinetobacter baumanii



Acinetobacter baumanii

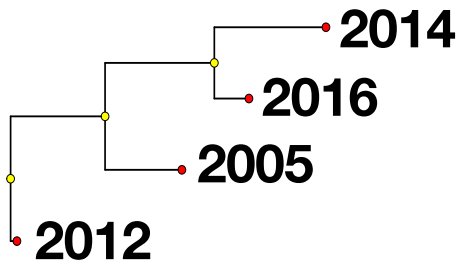
The date-randomisation test

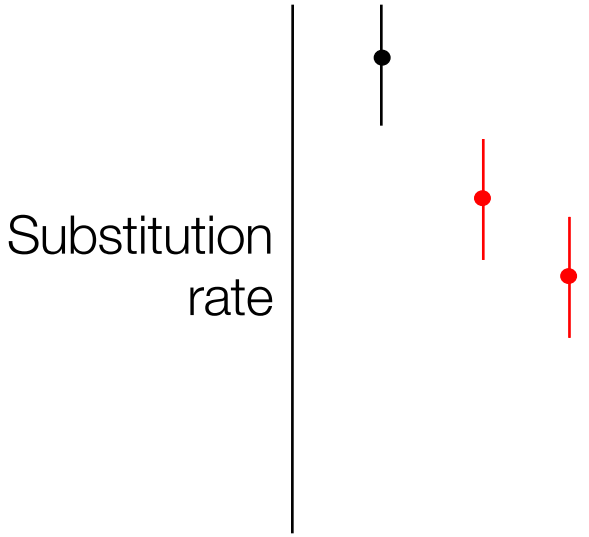




Duchêne et al. 2015 Mol Biol Evol

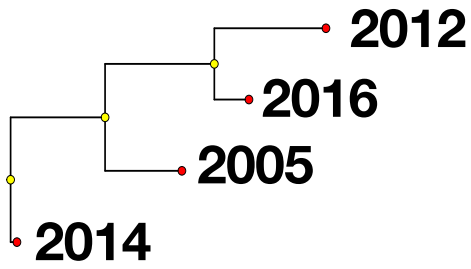
The date-randomisation test

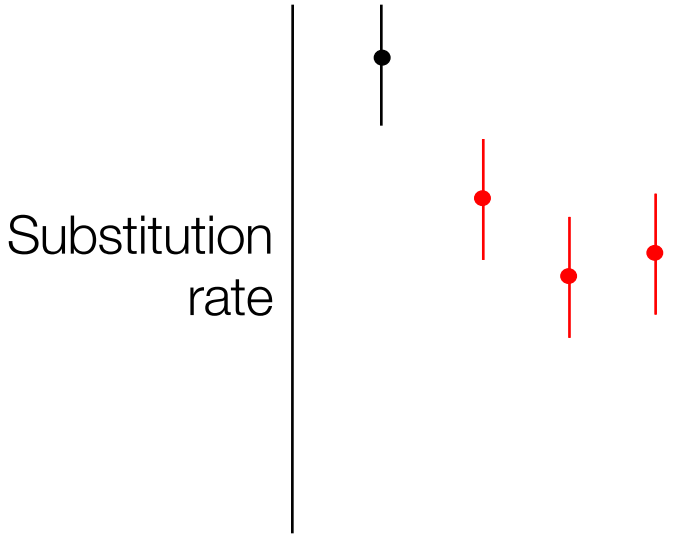




Duchêne et al. 2015 Mol Biol Evol

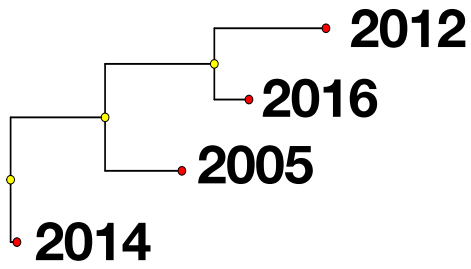
The date-randomisation test

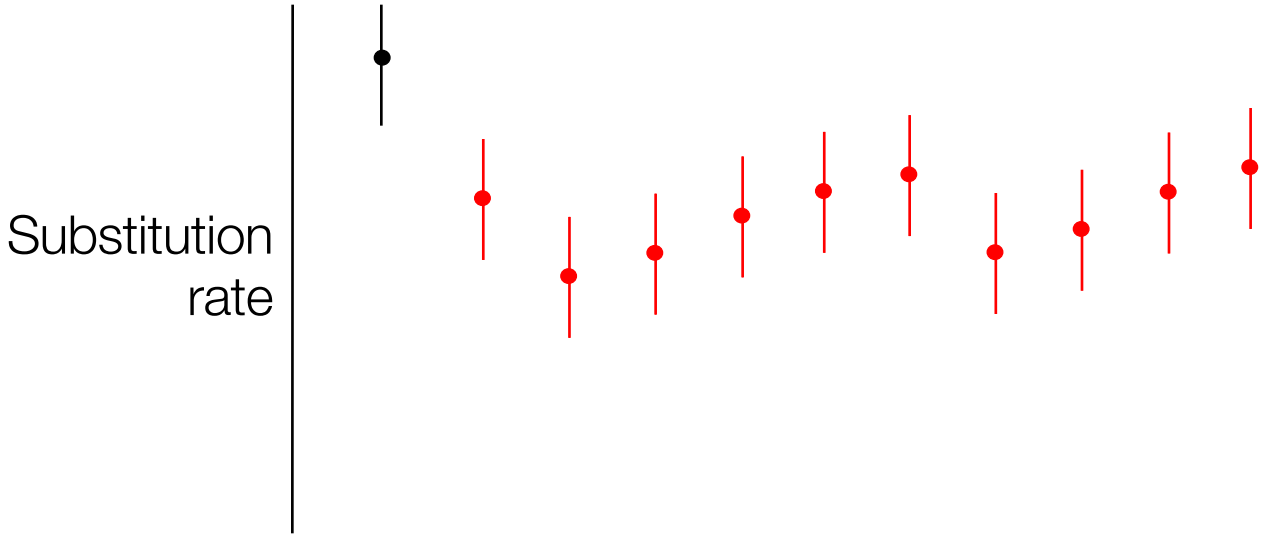




Duchêne et al. 2015 Mol Biol Evol

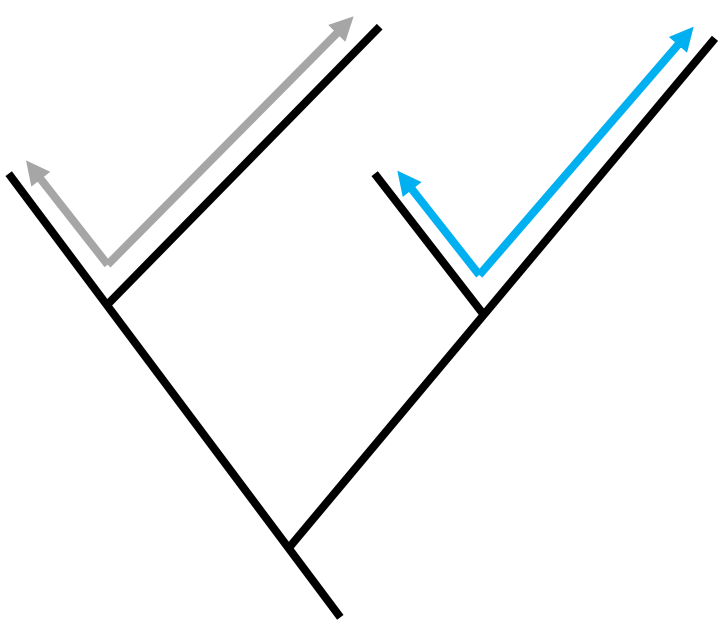
The date-randomisation test

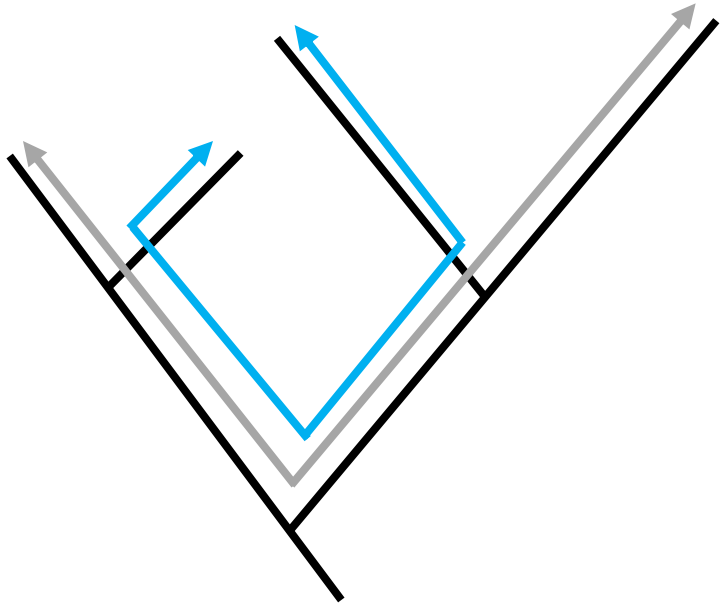




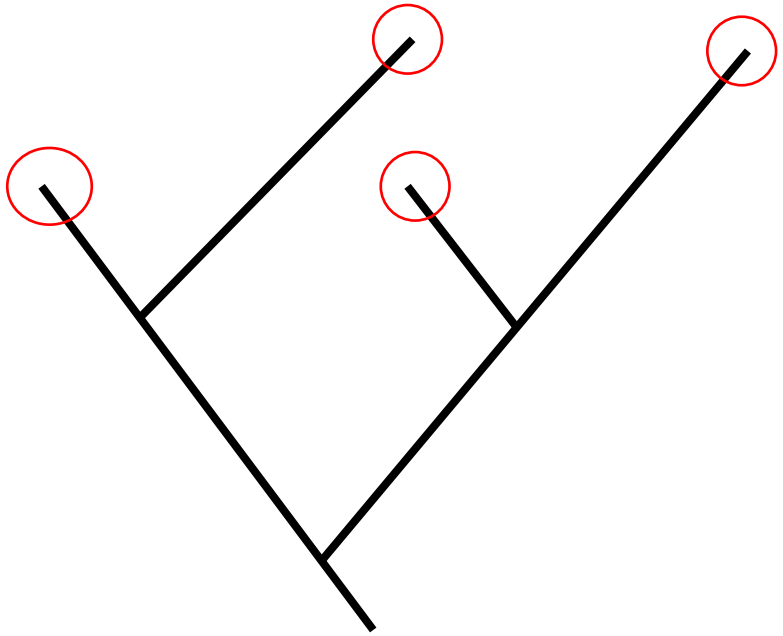
Duchêne et al. 2015 Mol Biol Evol

Phylogenetic and temporal clustering

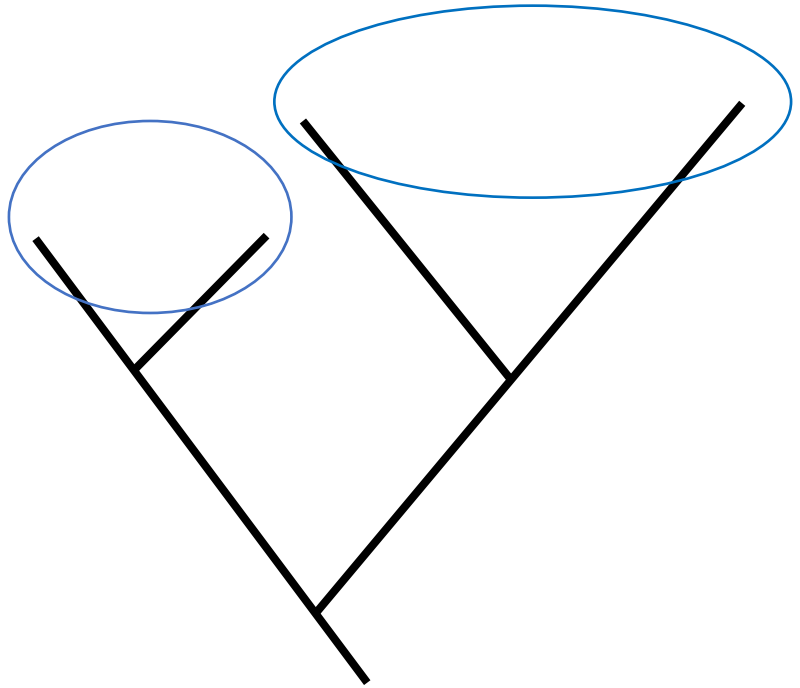




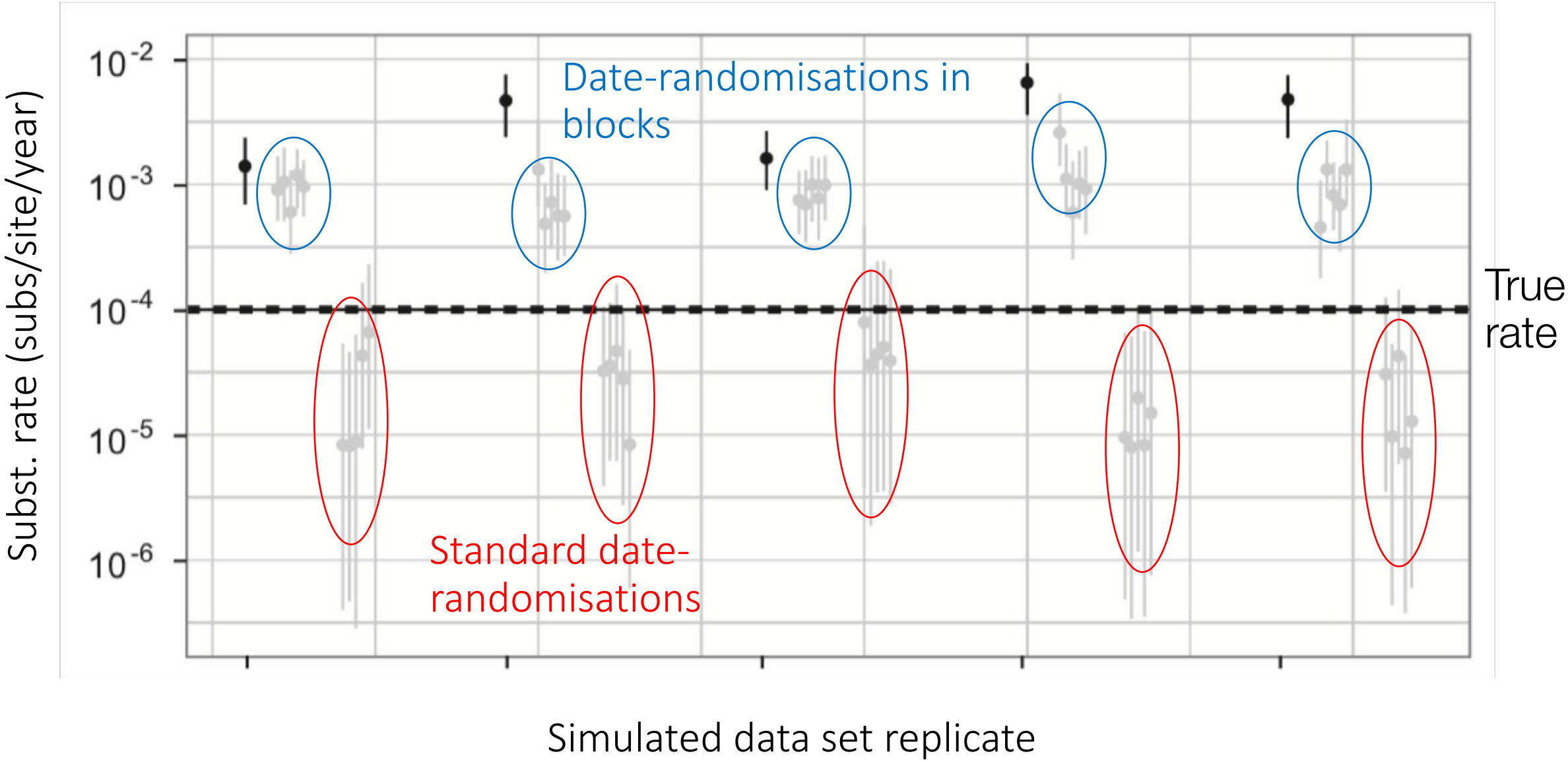
Date-randomisations in blocks



Murray *et al.* 2016 *Methods Ecol. Evol.*



Duchêne *et al.* 2015 *Mol. Biol. Evol.*


Subst. rate (subs/site/year)
10-2
10-3
10-4
10-5
10-6
Date-randomisations in blocks
Standard date-randomisations
True rate
Simulated data set replicate

Bayesian methods to assess temporal structure

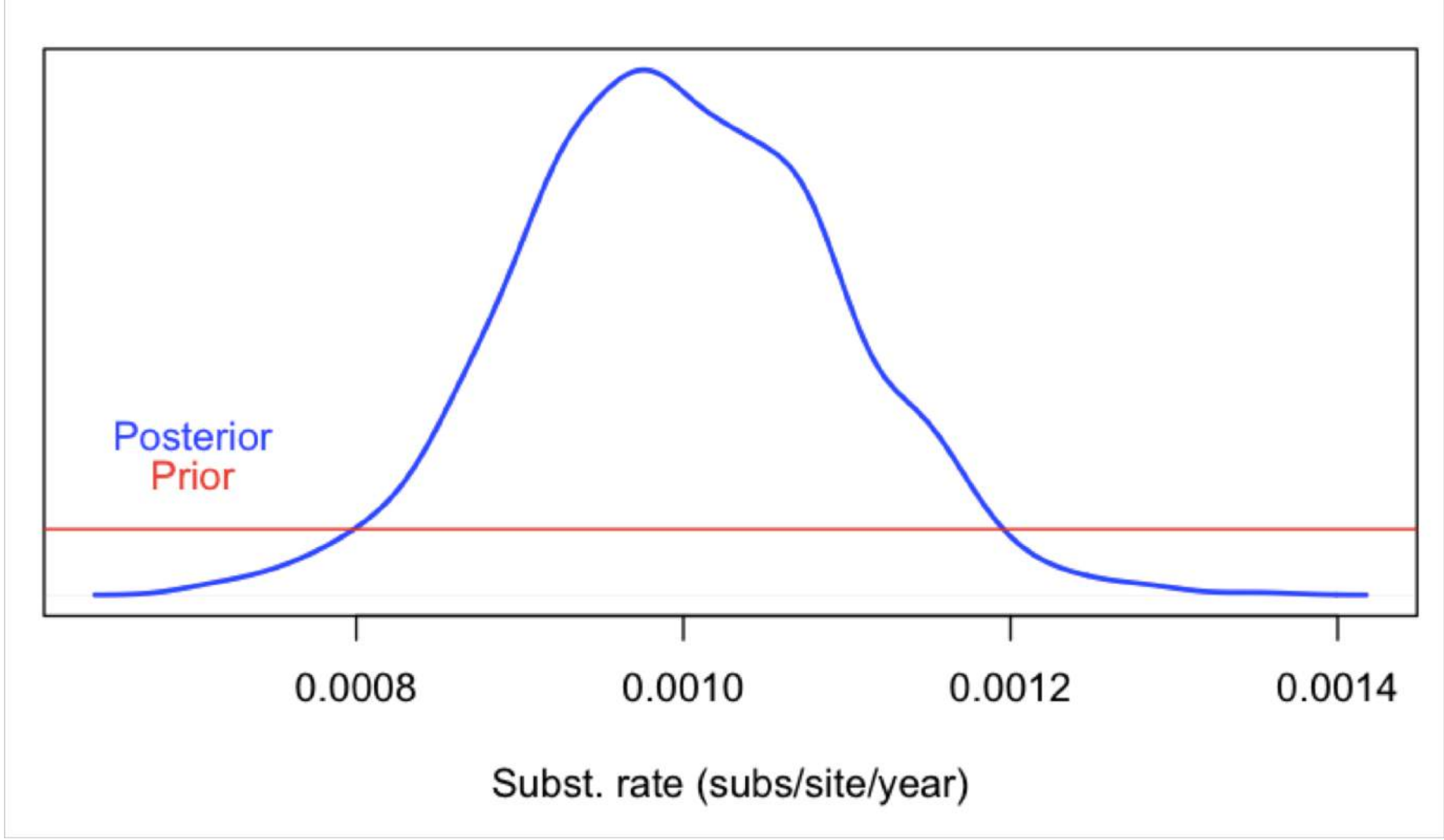
Duchene and Duchene. Submitted to:
The molecular evolutionary clock: theory and practice

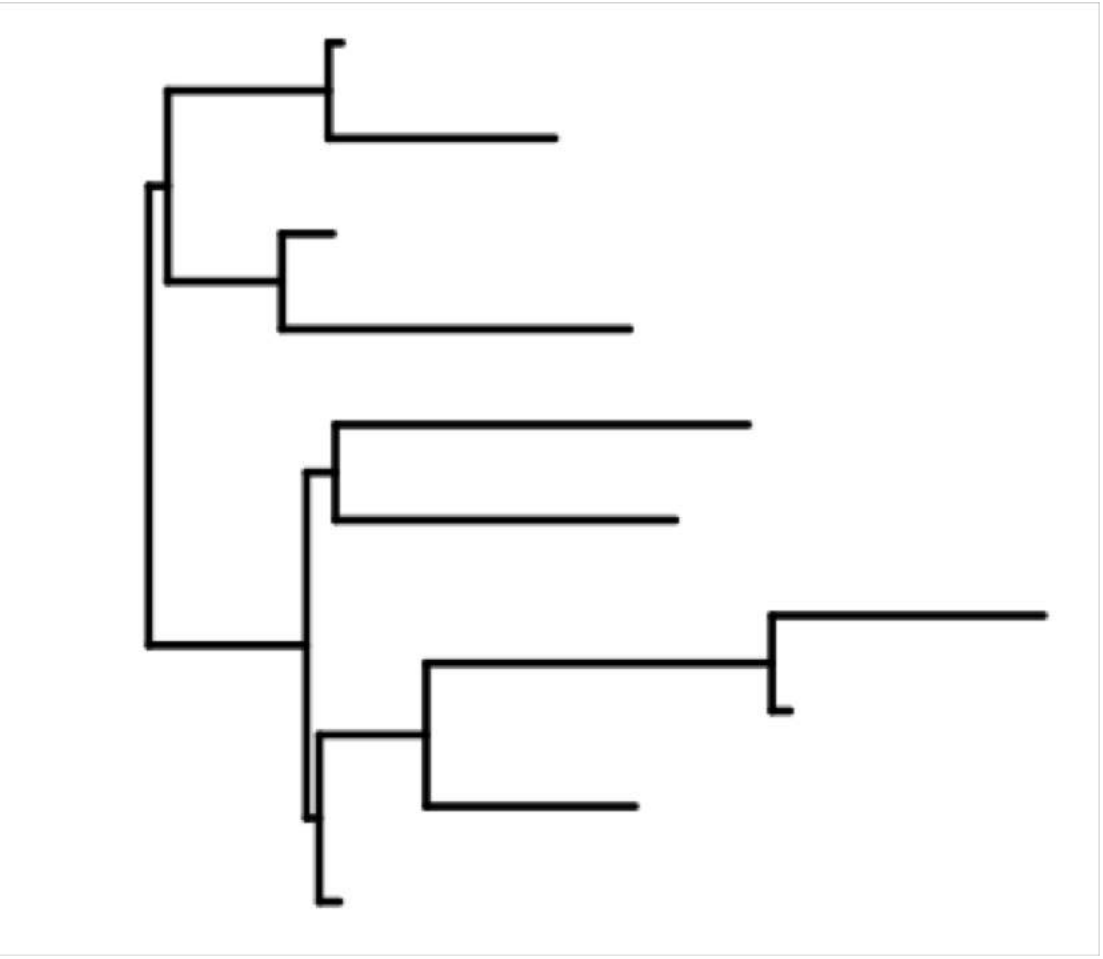
Comparing prior and posterior distributions

Alignment | Time tree | Branching model (can be an epi model) | Substitution model | Clock model

Sampling from the prior means that we do not calculate the phylogenetic likelihood.

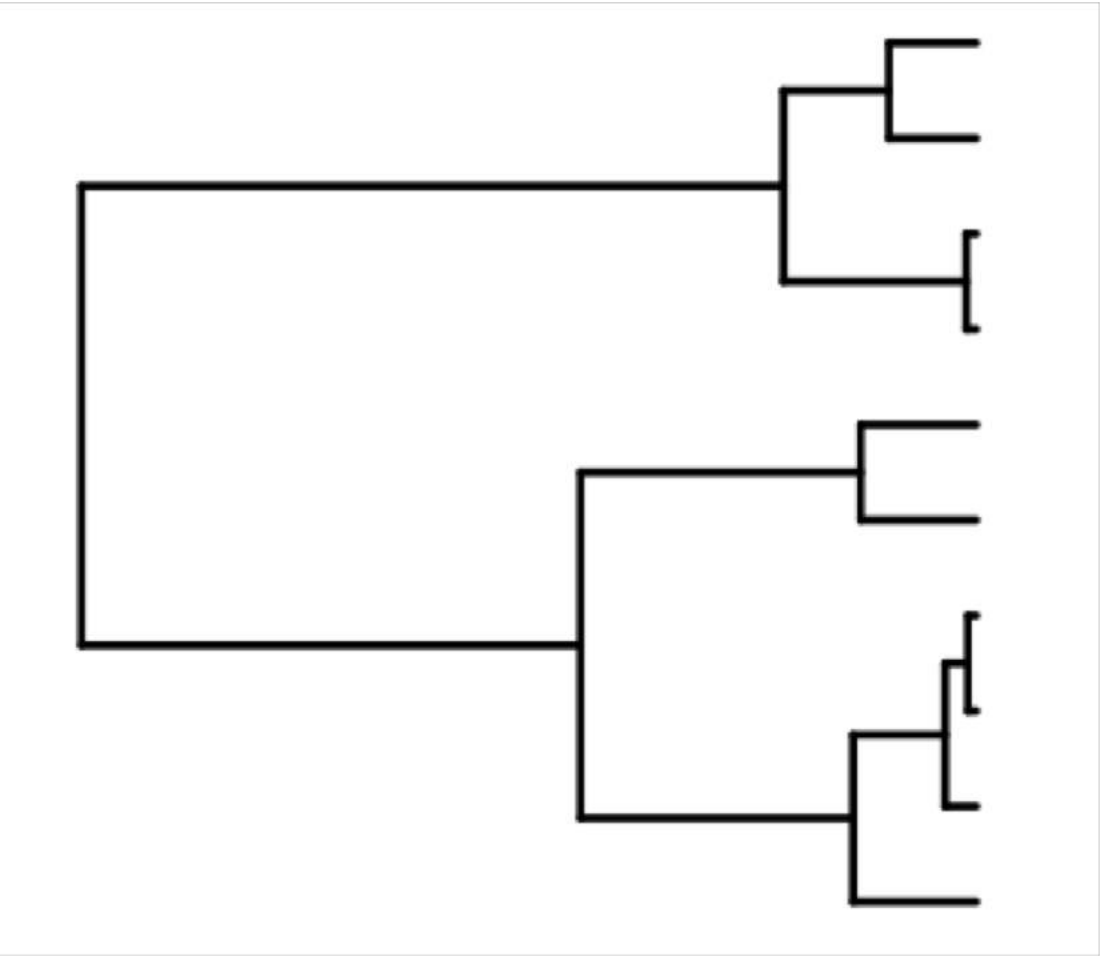
Comparing prior and posterior distributions





With no sequence data the root height and rate are driven by the prior.

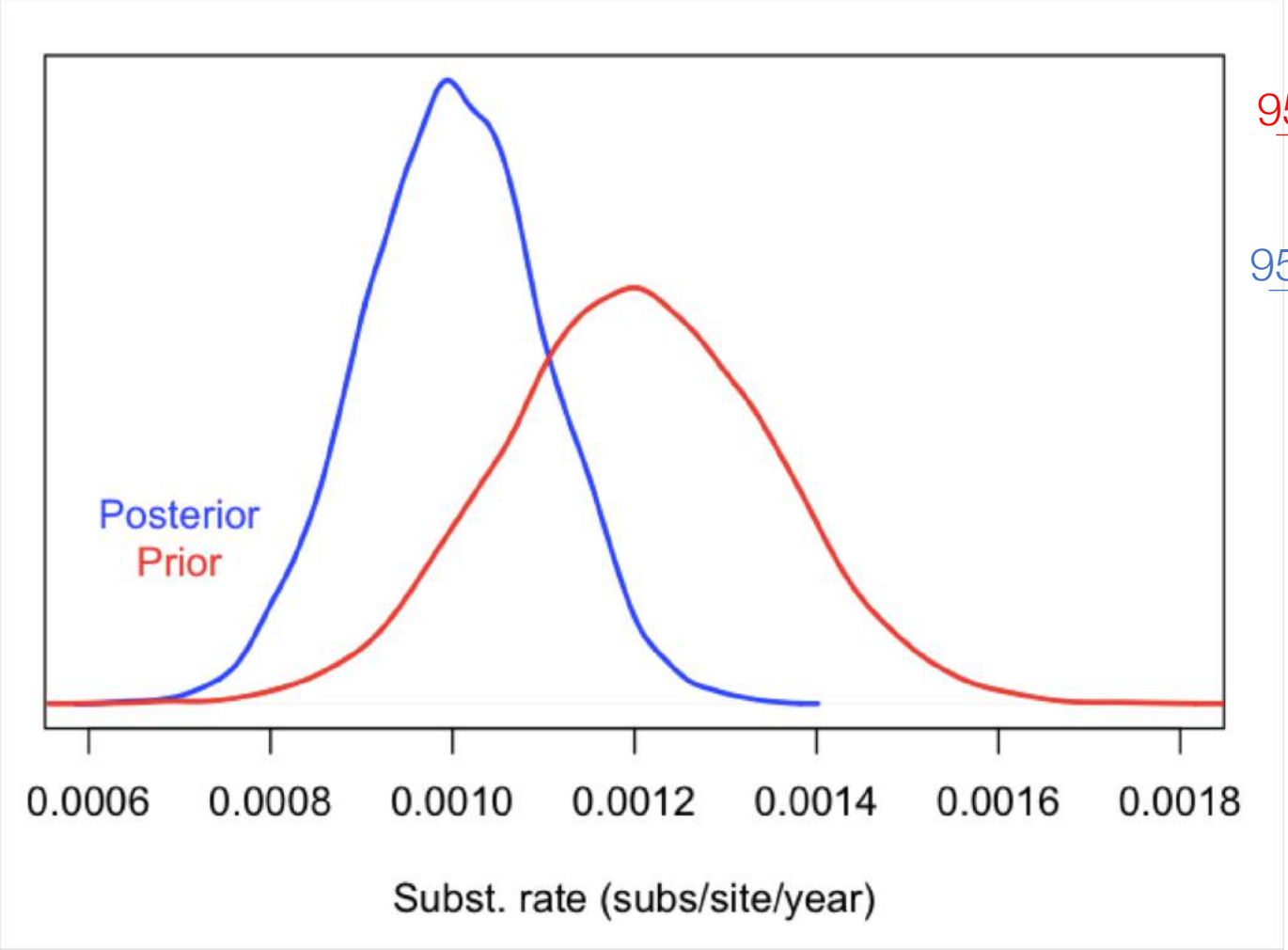
If the data have temporal structure the root height and rate should have a posterior that is much more informative than the prior.



If the sampling times are uninformative there is effectively no calibration information.

The root height should be driven by the prior.

Quantifying information content



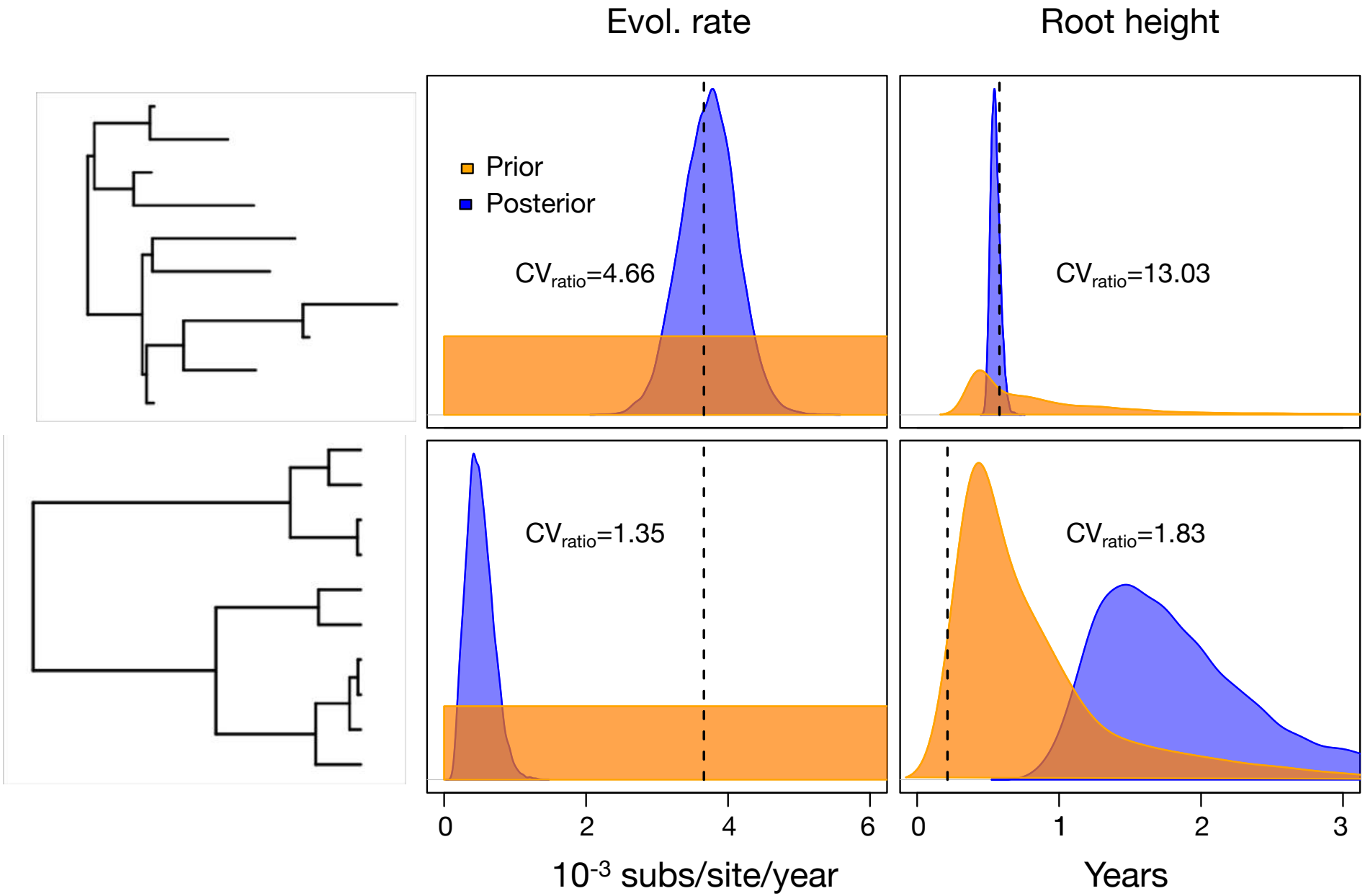
$$\frac{\text{95\% quantile width}}{\text{Mean value}} = CV_{prior}$$

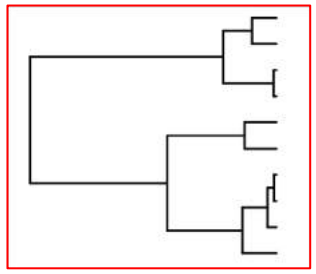
$$\frac{\text{95\% quantile width}}{\text{Mean value}} = CV_{posterior}$$

$$\frac{CV_{prior}}{CV_{posterior}} = CV_{ratio}$$

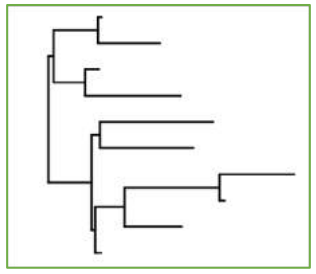
$CV_{ratio} > 1$ Posterior is more informative than the prior.

$CV_{ratio} = 1$ Posterior and prior are equally informative. i.e. estimates are driven by the prior.

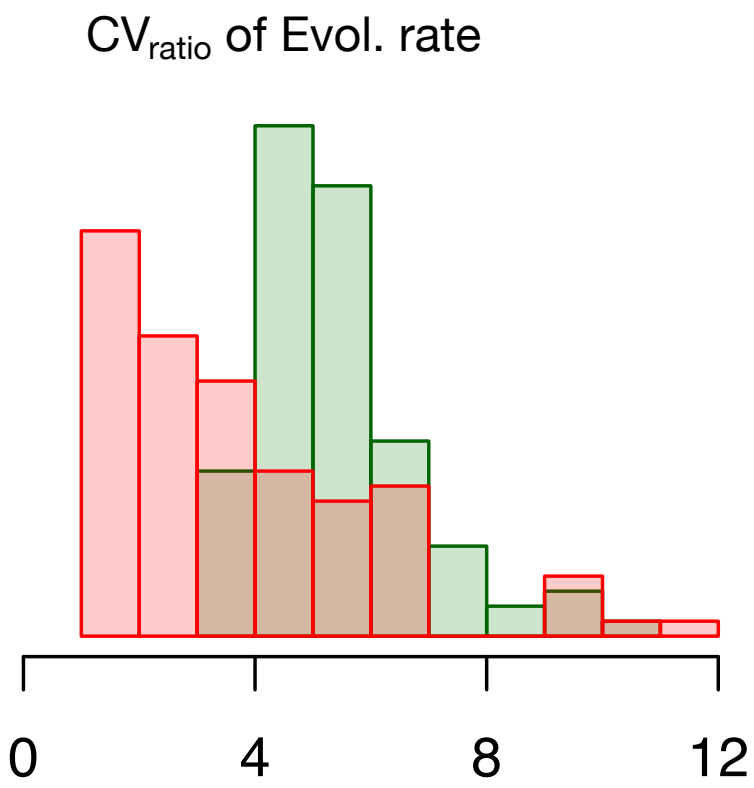

Evol. rate
Root height
Prior
Posterior
CV$_{ratio}$=4.66
CV$_{ratio}$=13.03
CV$_{ratio}$=1.35
CV$_{ratio}$=1.83
0
2
4
6
0
1
2
3
10^{-3} subs/site/year
Years

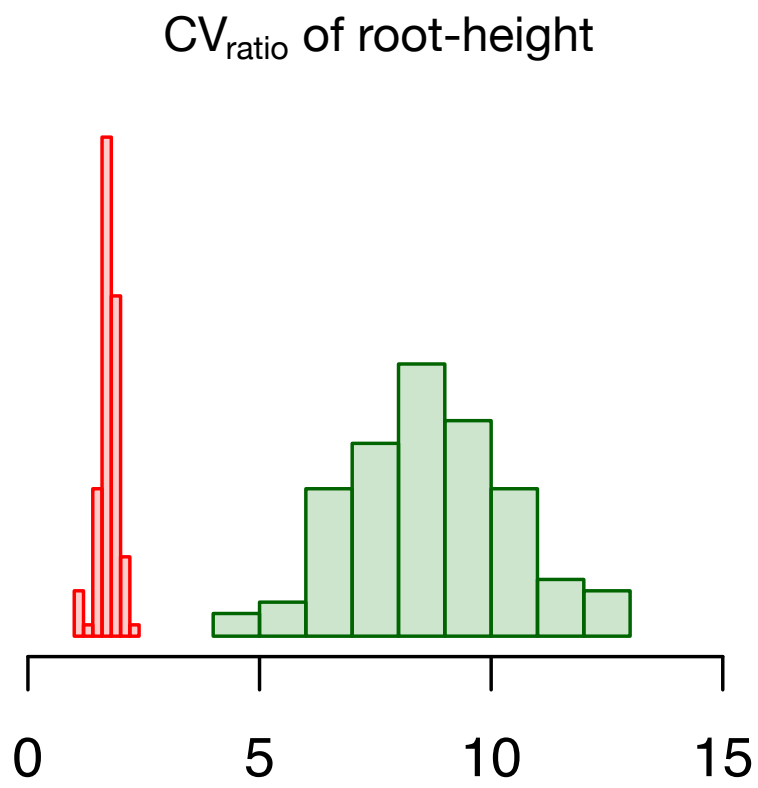


No temporal structure

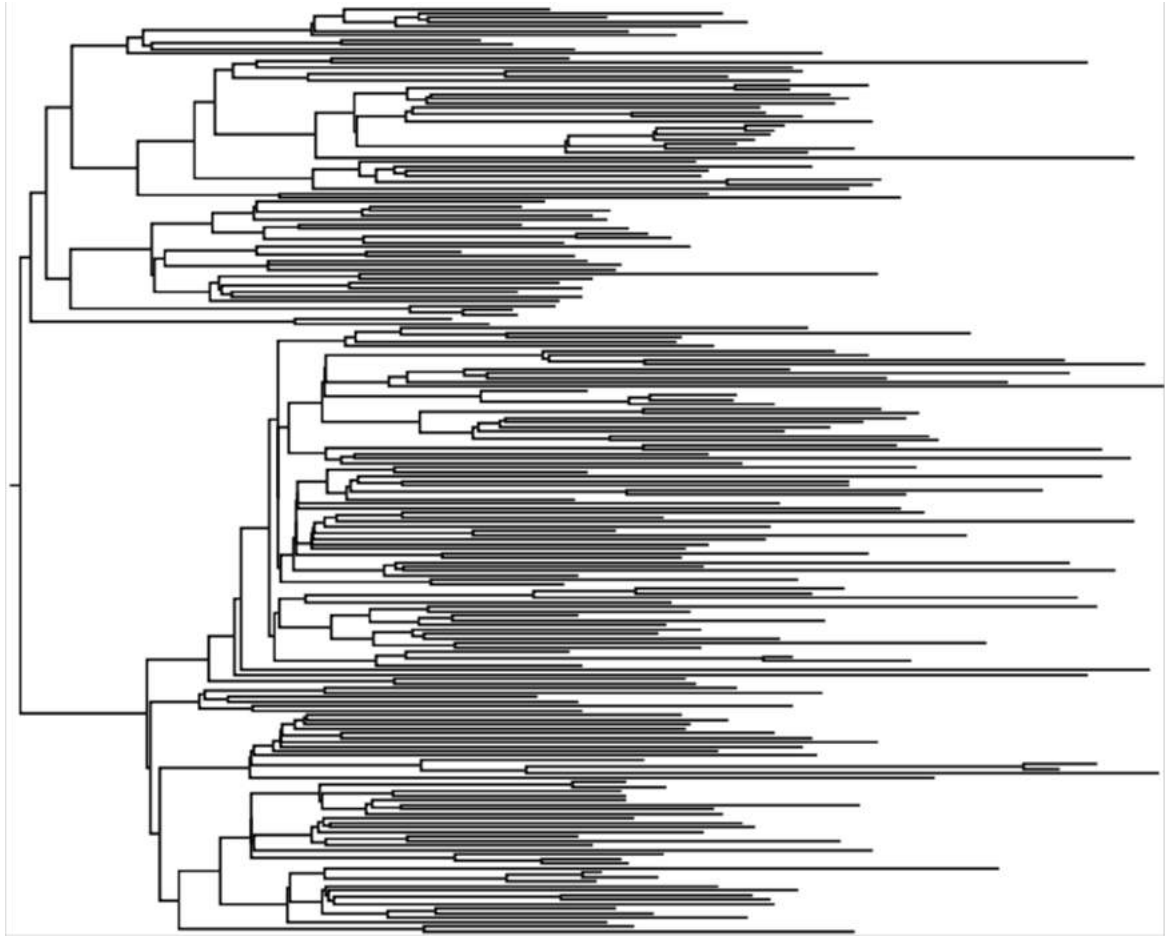


Temporal structure

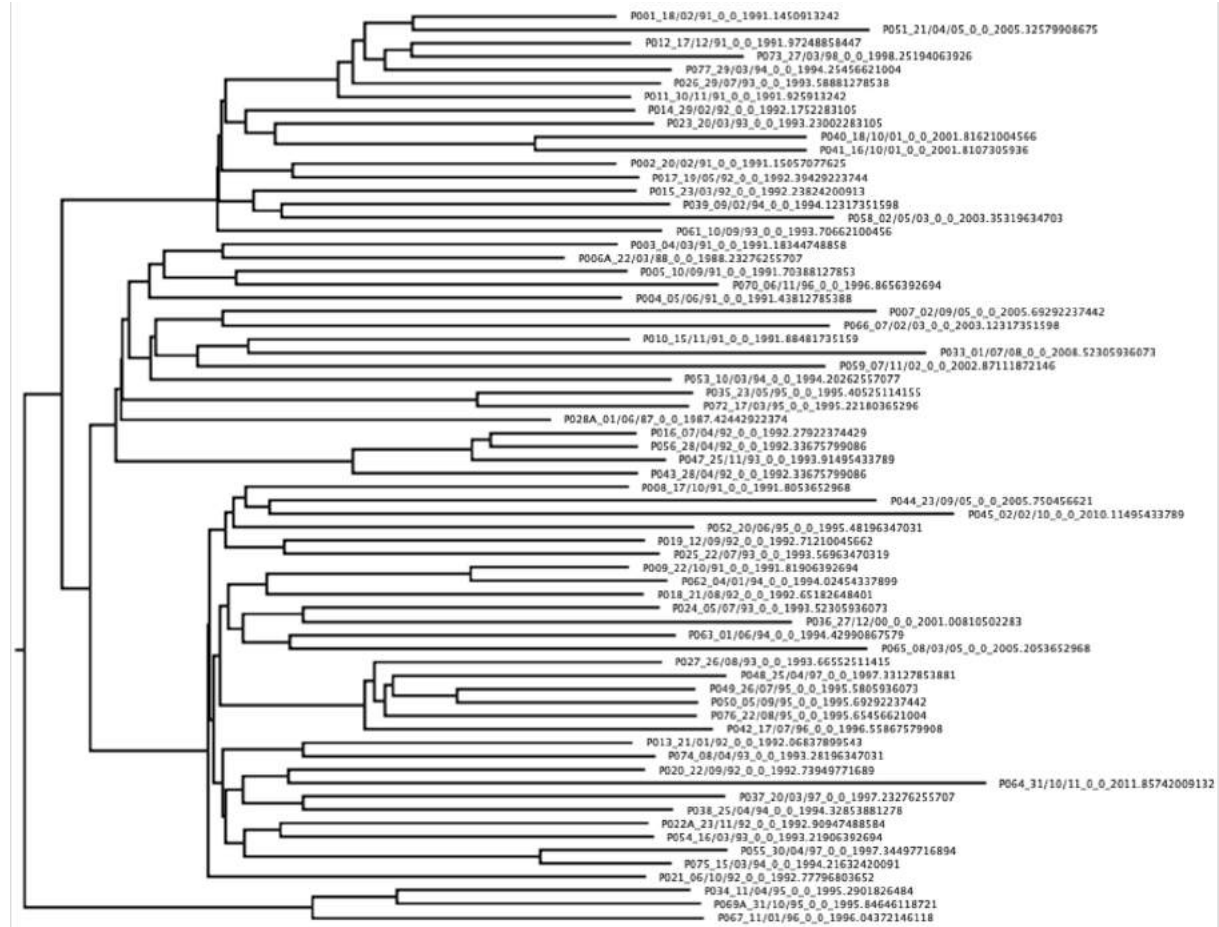


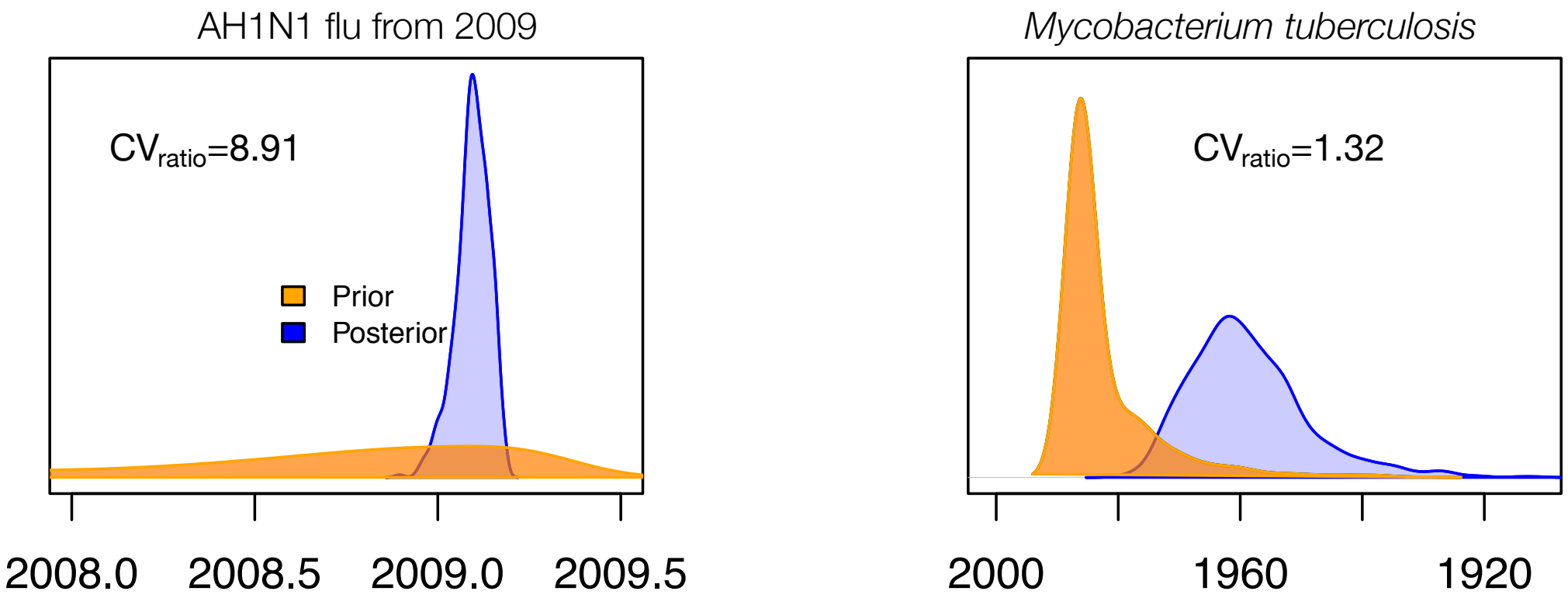


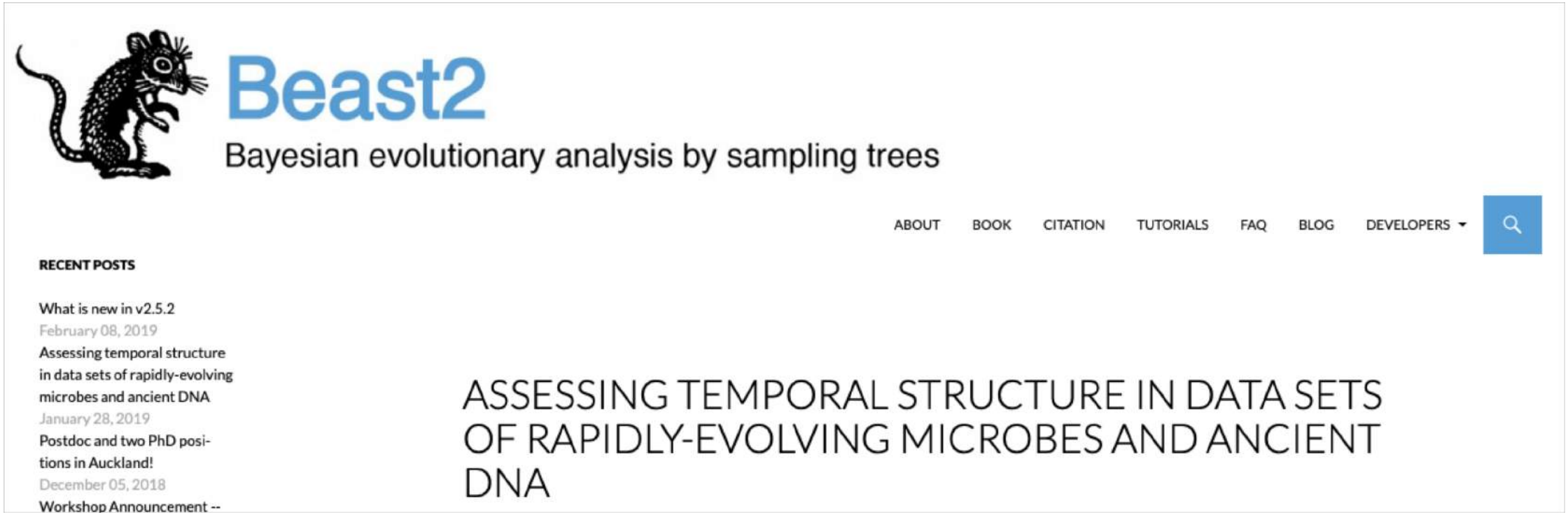
AH1N1 flu from 2009



Mycobacterium tuberculosis




AH1N1 flu from 2009
CV ratio=8.91
Prior
Posterior
2008.0
2008.5
2009.0
2009.5
Mycobacterium tuberculosis
CV ratio=1.32
2000
1960
1920
Age of root-node



Duchene and Duchene. Submitted.

Taming the BEAST 

news workshops tutorials contribute

Molecular dating using heterochronous data and substitution model averaging

Estimating the time of origin of the 2009 H1N1 pandemic in North America

by Sebastian Duchene , Tim Vaughan and Veronika Boskova

Tutorial

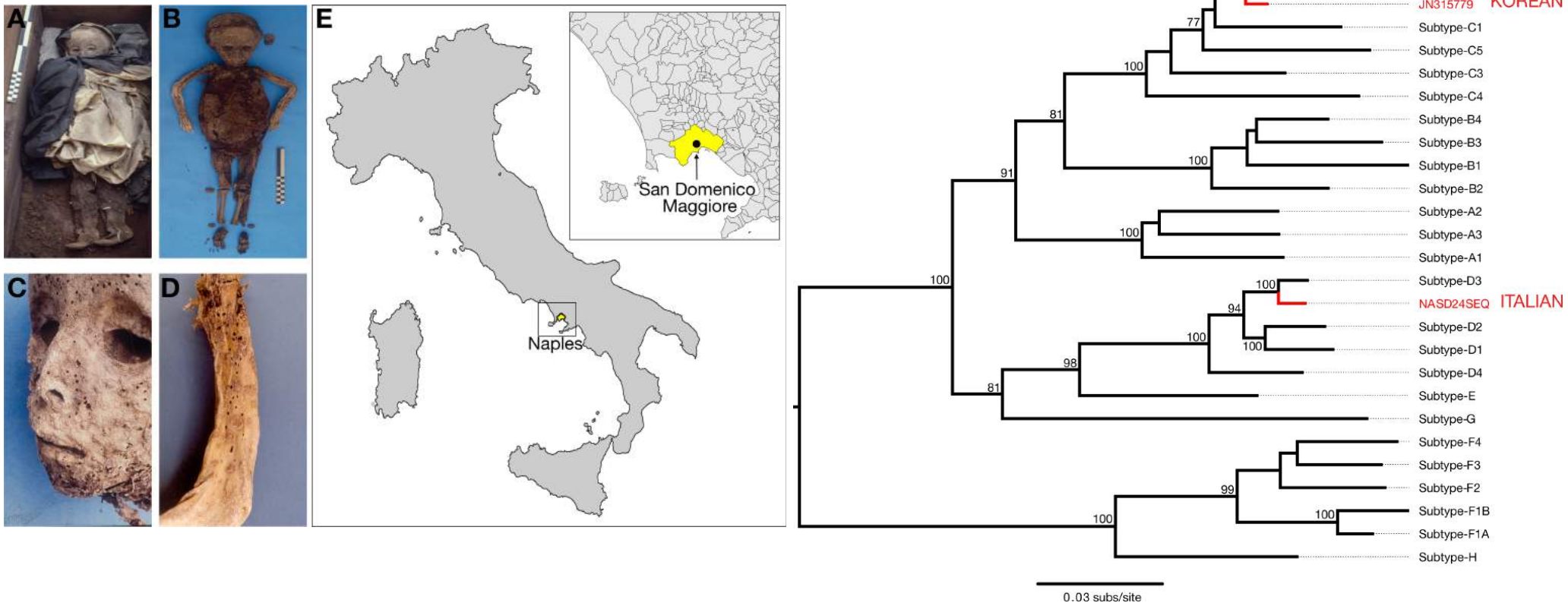
Github repository

Background

A case study in HBV (ancient DNA)

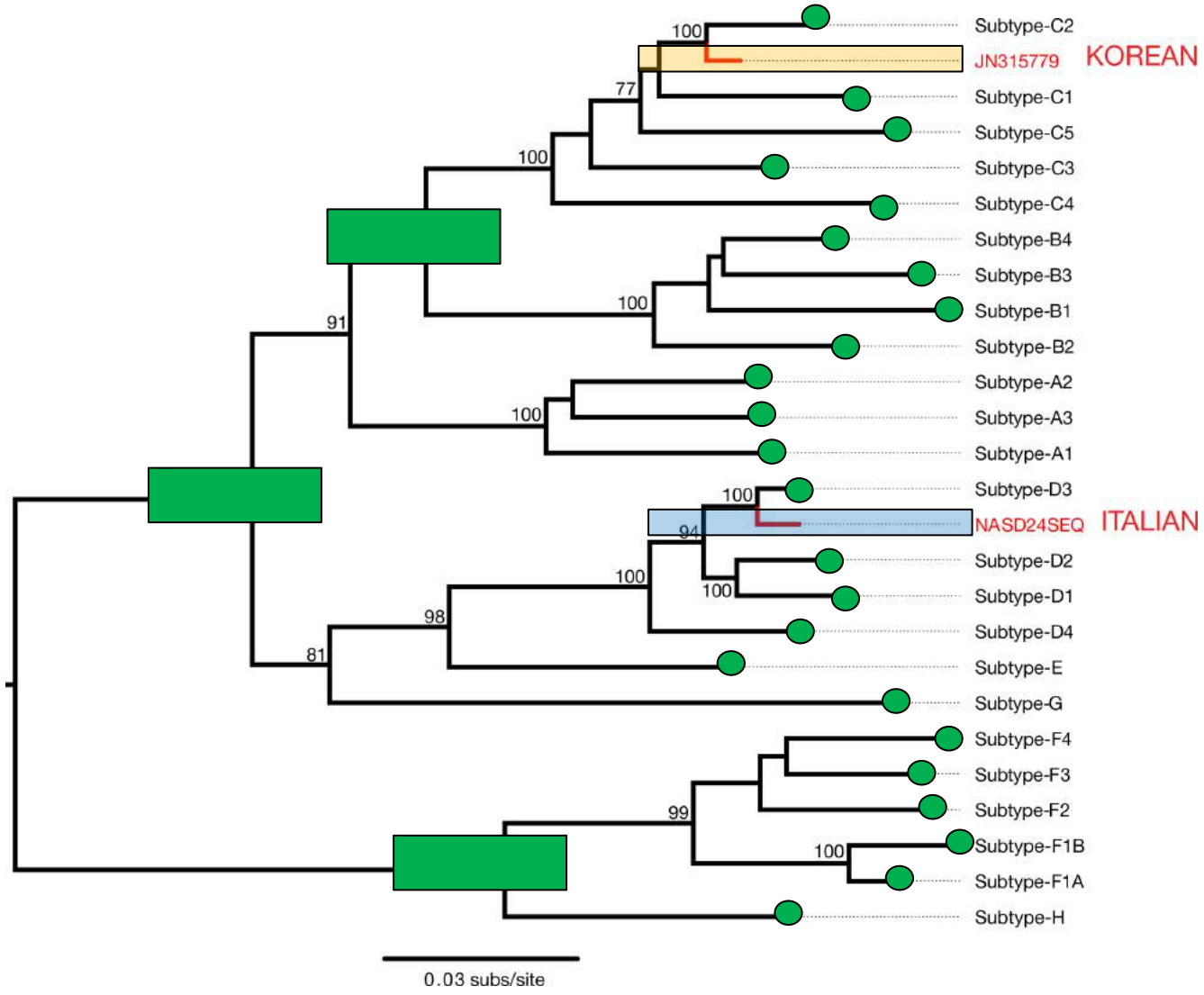
Comparing the prior and posterior -tips

Hepatitis B Virus paradox

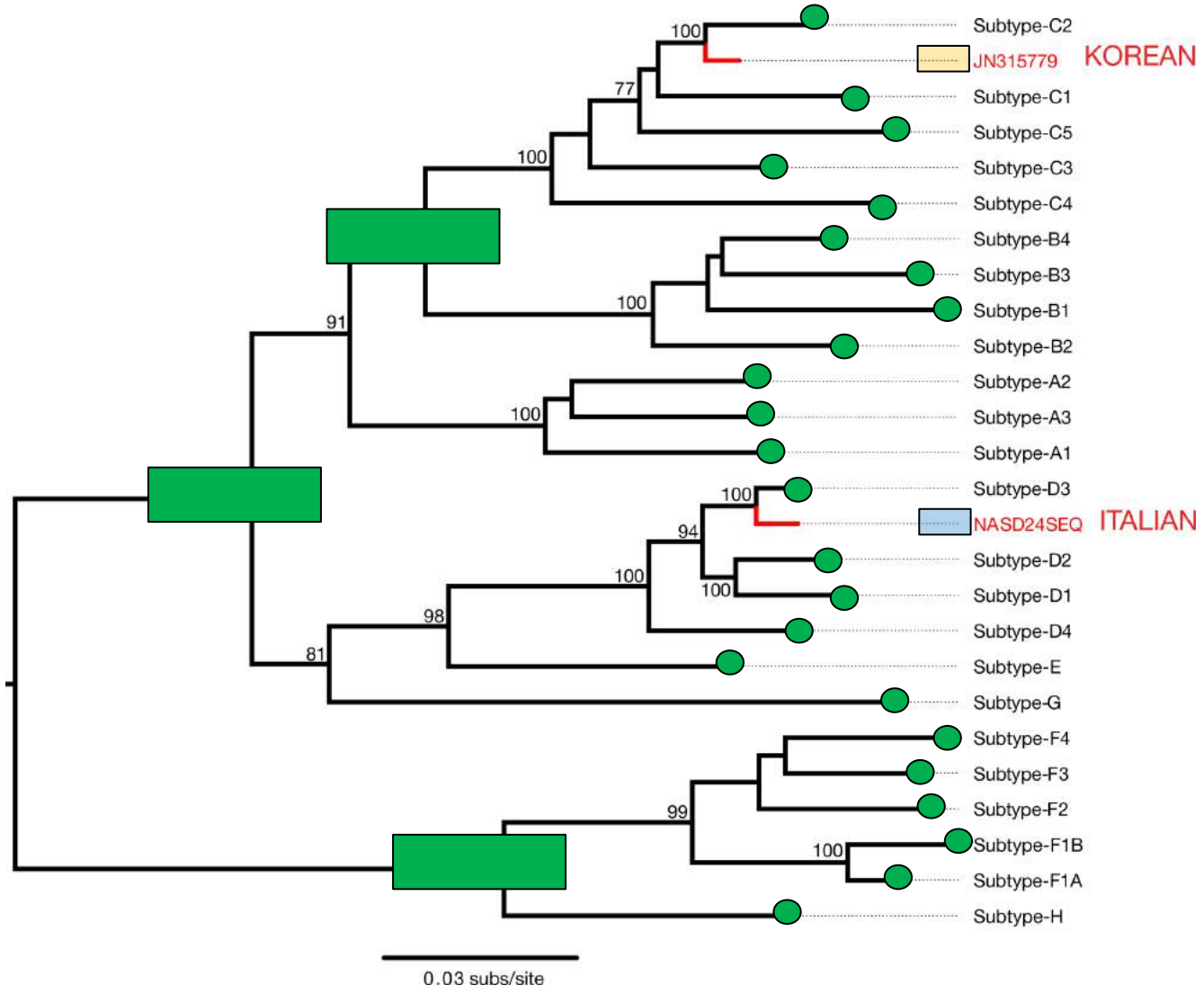


Patterson Ross *et al.* 2018 *PLoS Pathogens*

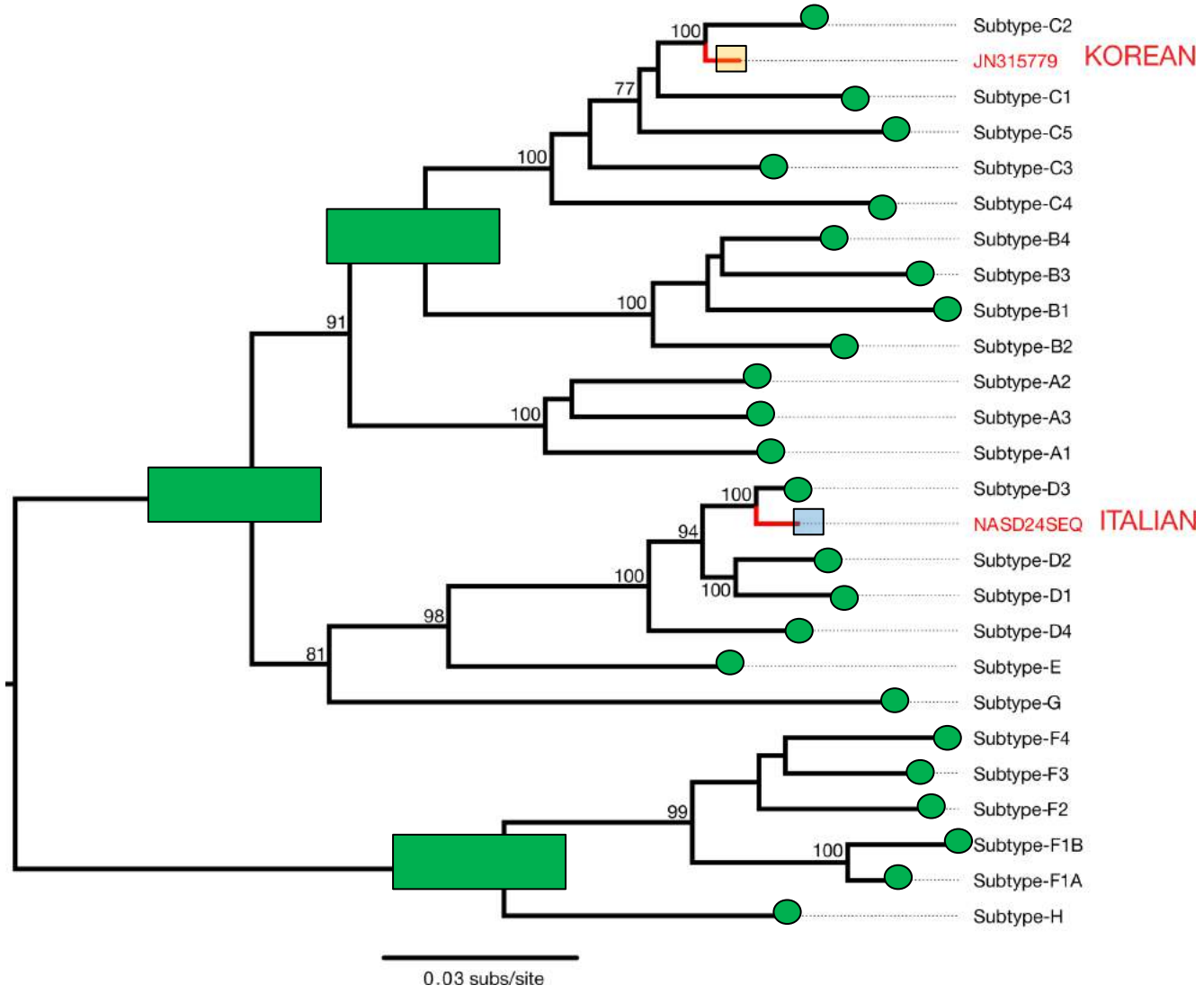
Comparing the prior and posterior -tips

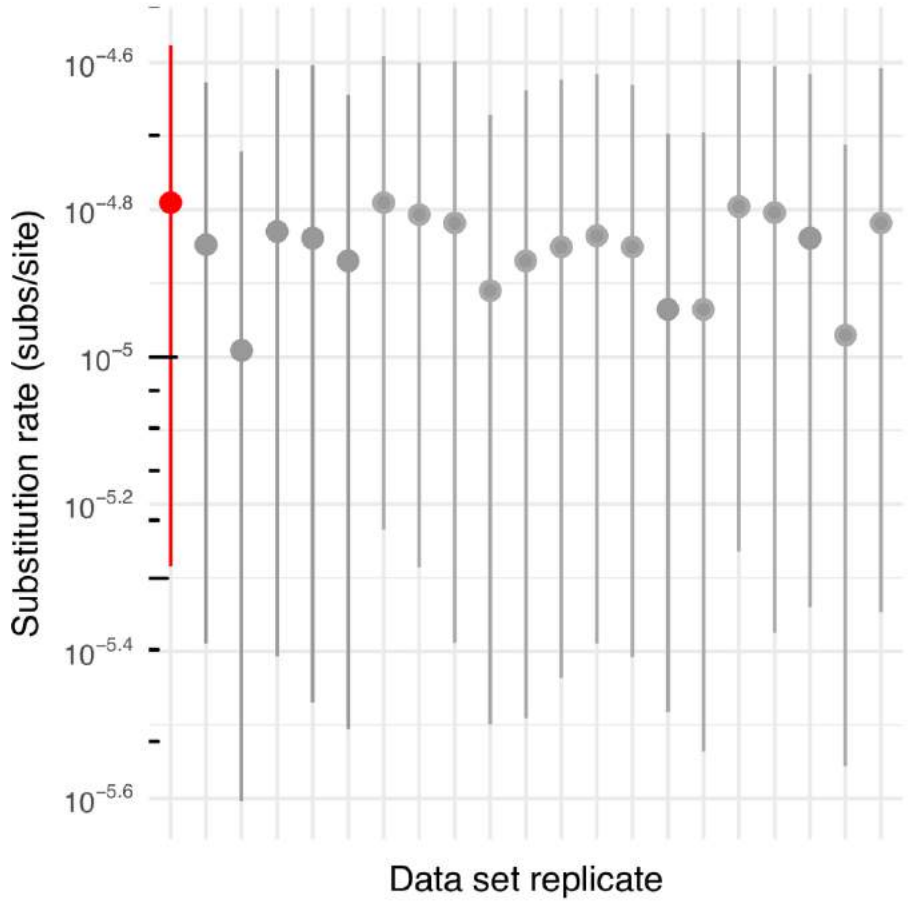


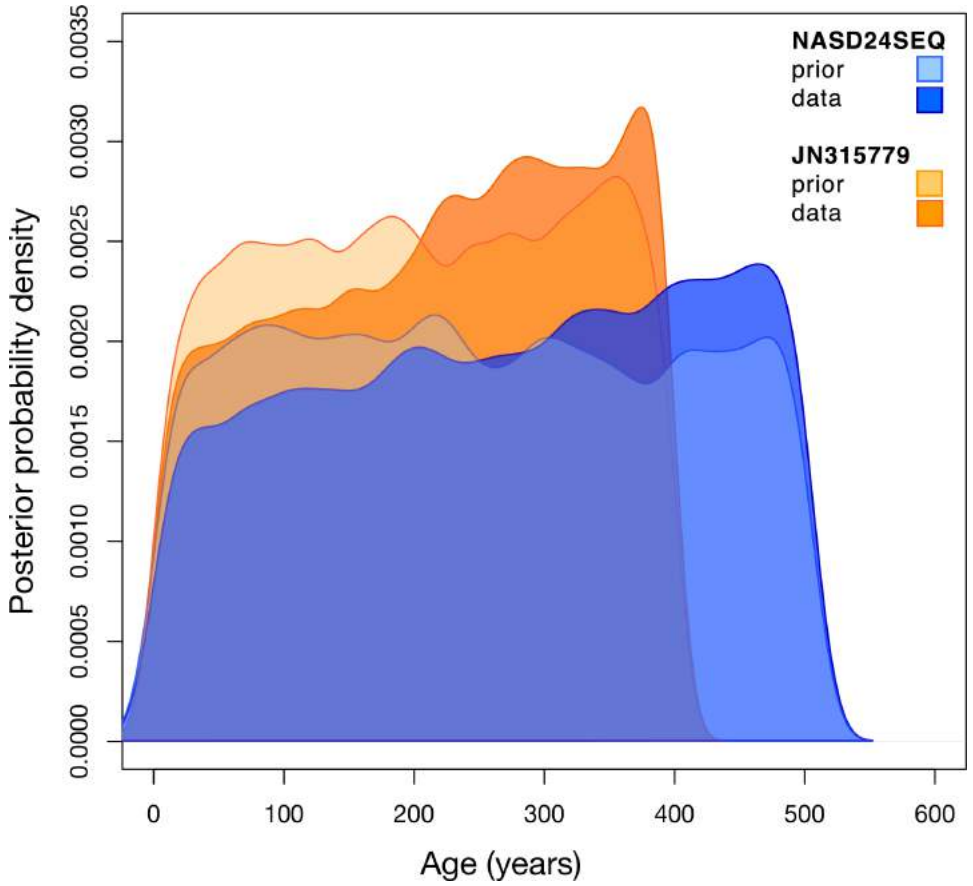
Comparing the prior and posterior -tips



Comparing the prior and posterior -tips



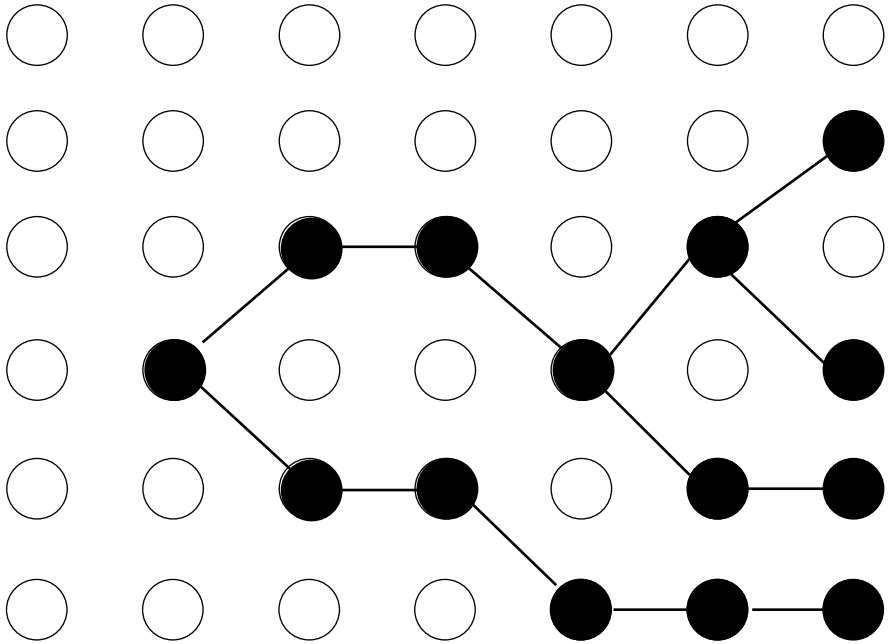


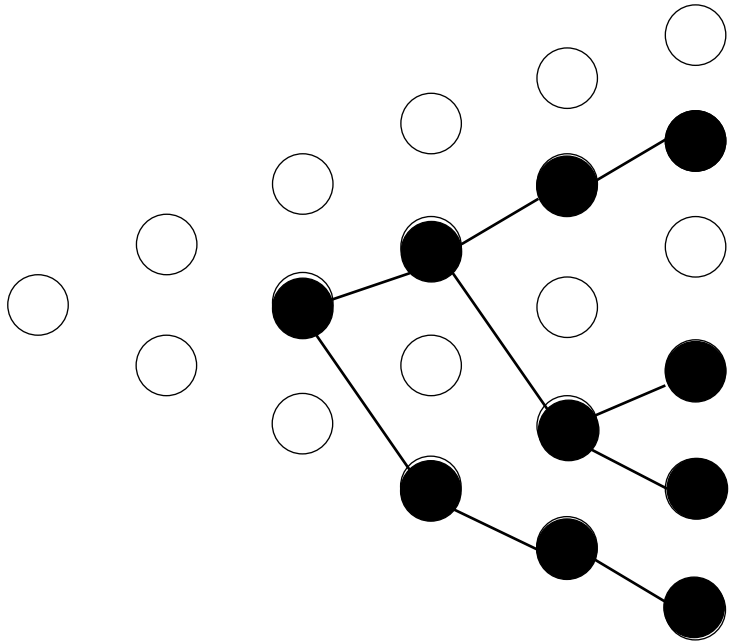


Patterson Ross *et al.* 2018 *PLoS Pathogens*

Tree priors and tip-dating

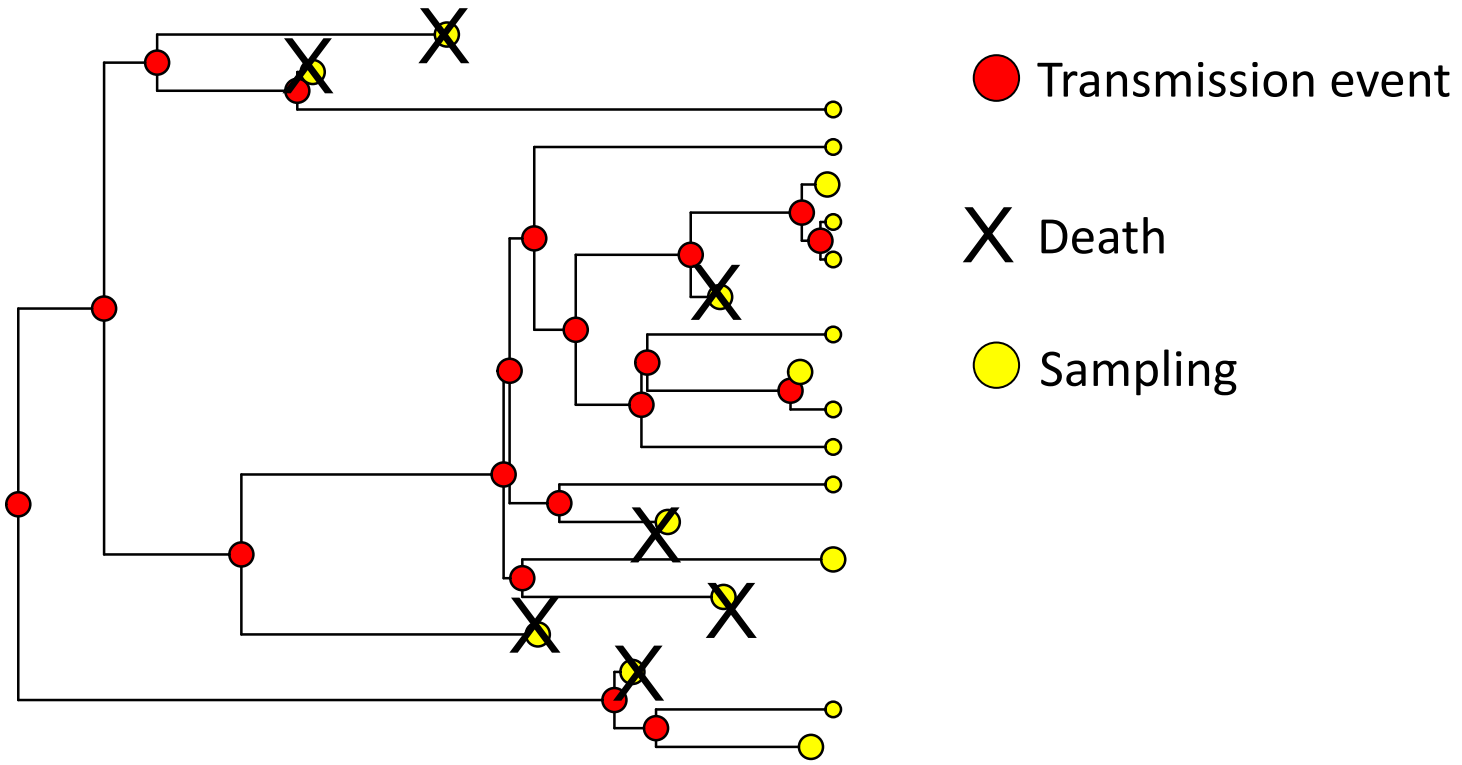
The coalescent





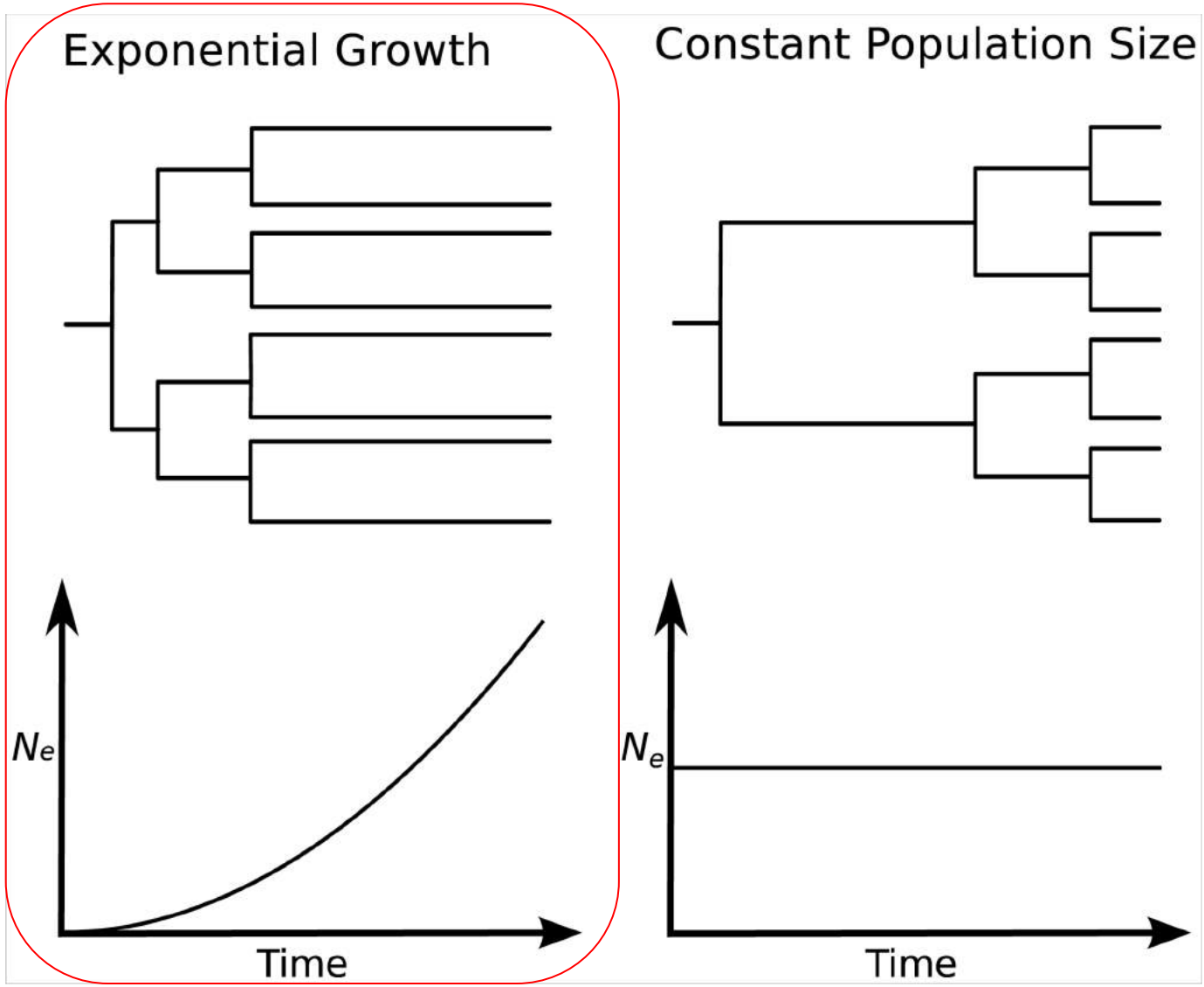
Backwards-in-time process

The birth-death sampling through time



Forward-in-time process

Both, the birth-death
with sampling and the
exponential
coalescent model
exponential
growth!



From: Volz et al. 2013

Epidemiological parameters

- The coalescent and birth-death have different parameterisations.

Coalescent exponential

r: growth rate
Φ: effective population size
I(0): infected population size at present

Birth-death

R_0: basic reproductive number
δ: become uninfectious rate
p: sampling probability

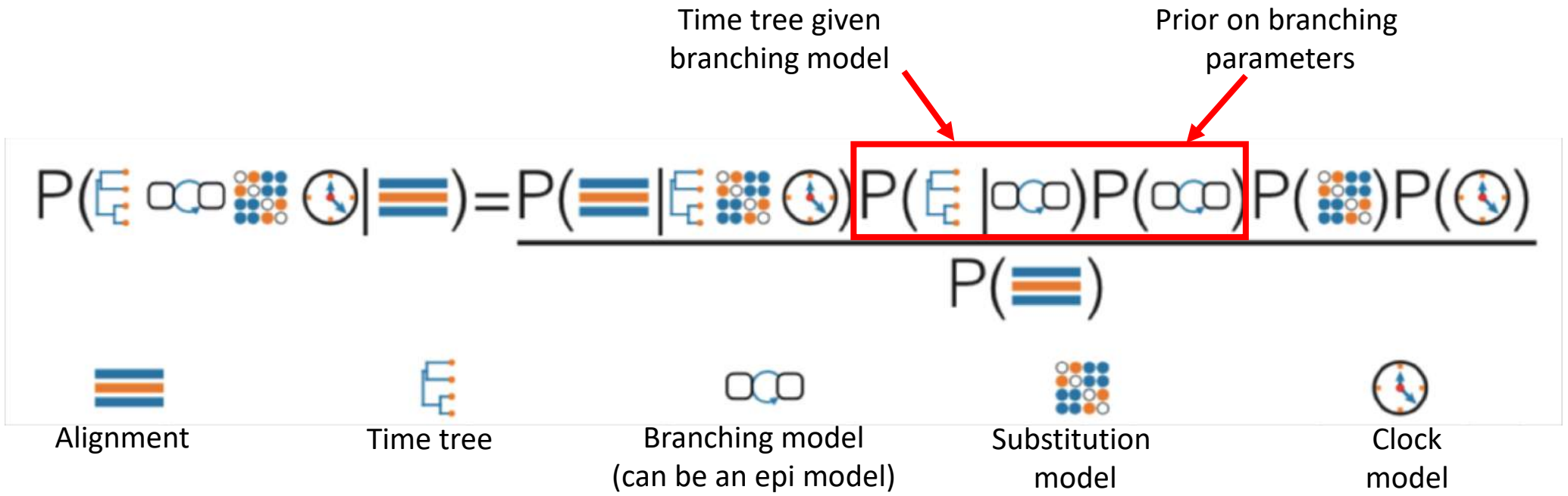
$1 / \delta$ = duration of infection

$\Phi = I(0) / 2\lambda$
$r = \lambda - \delta$

$R_0 = (r + \delta) / \delta$

$R_0 = \lambda / \delta$
$\delta = \mu + \psi$
$p = \psi / (\mu + \psi)$

λ: transmission rate
ψ: sampling rate
μ: death rate



In the birth-death and the exponential coalescent  depends on $\lambda - \delta$ and $\lambda\delta\rho$ so we usually need a prior on any one parameter to calculate any individual parameter.

From: Du Plessis and Stadler 2015

Tree prior $P(\text{tree} \mid \text{model})$

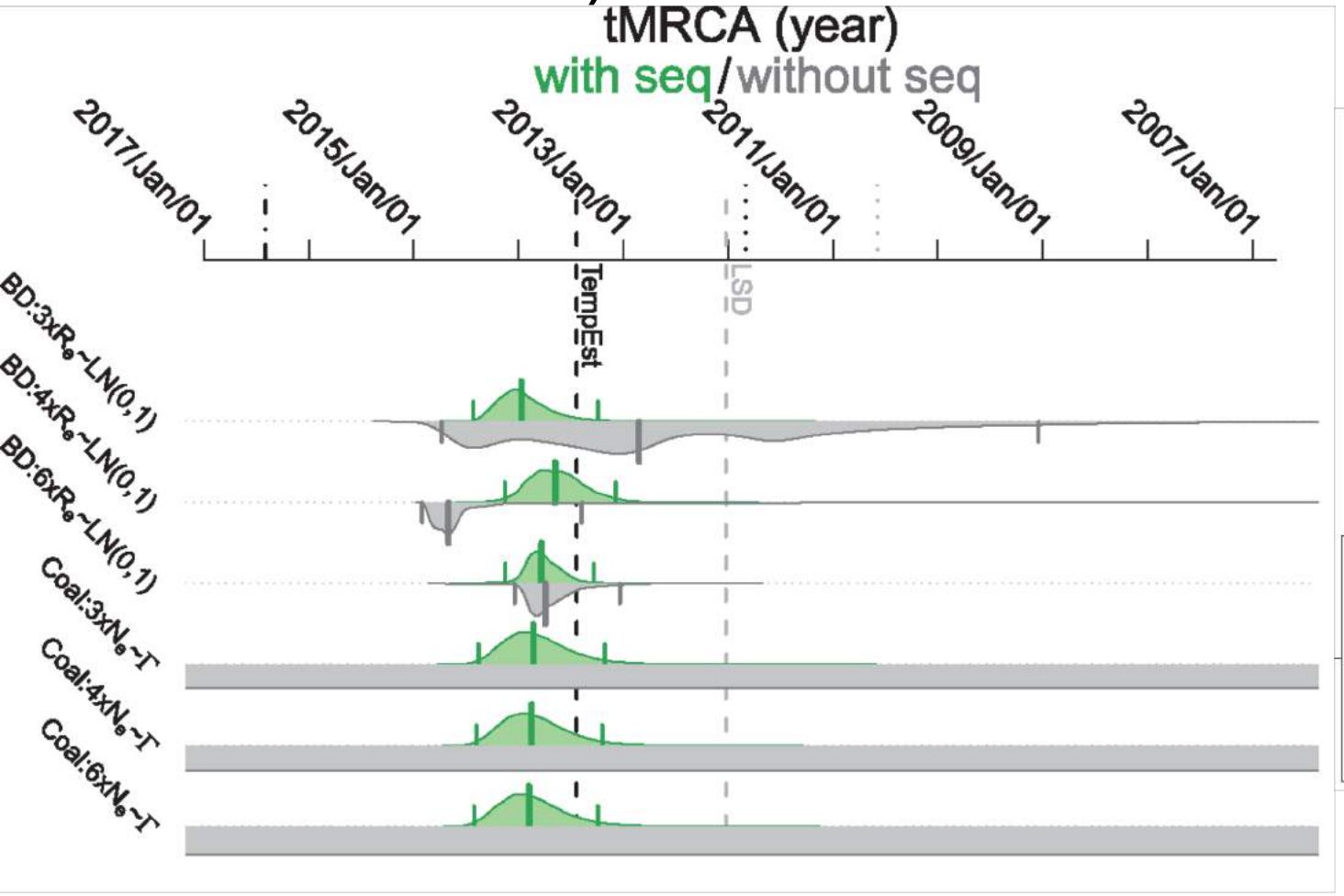
$P_{CoalSamp}(\text{tree} \mid par = \Phi, r)$

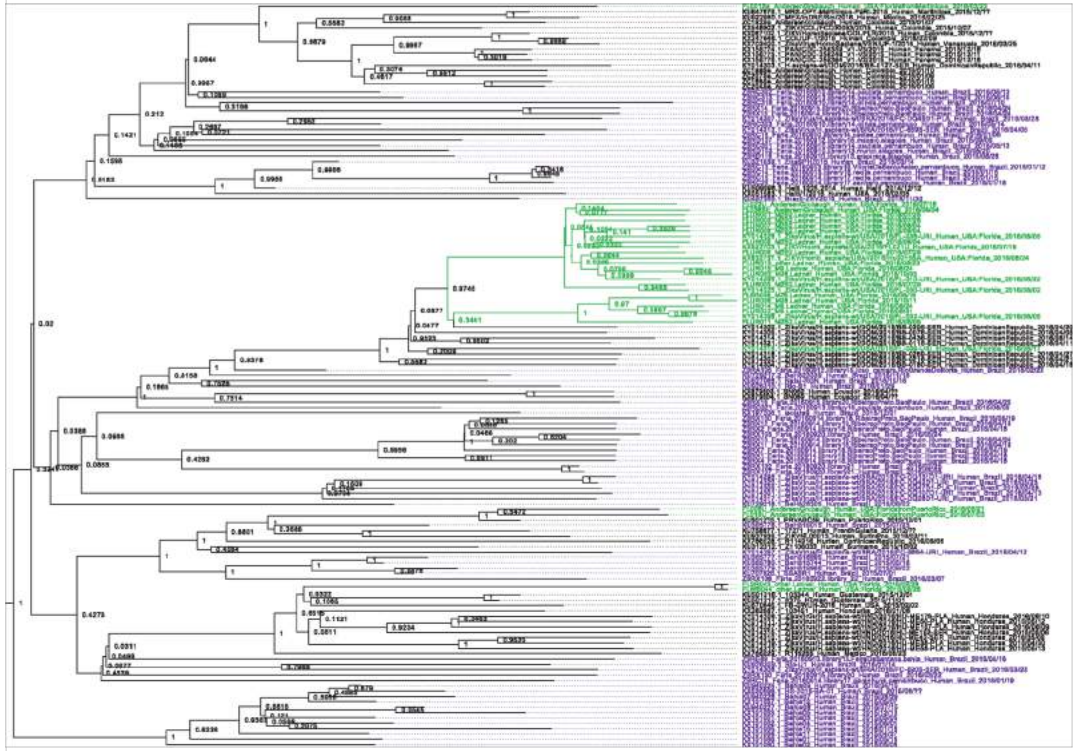
$P_{birth\text{-}death}(\text{tree}, \text{sampling times}, N \mid par = R_0, \delta, p)$

In the birth-death the number of samples and their ages are informative (treated similar to data).

Boskova et al. 2018

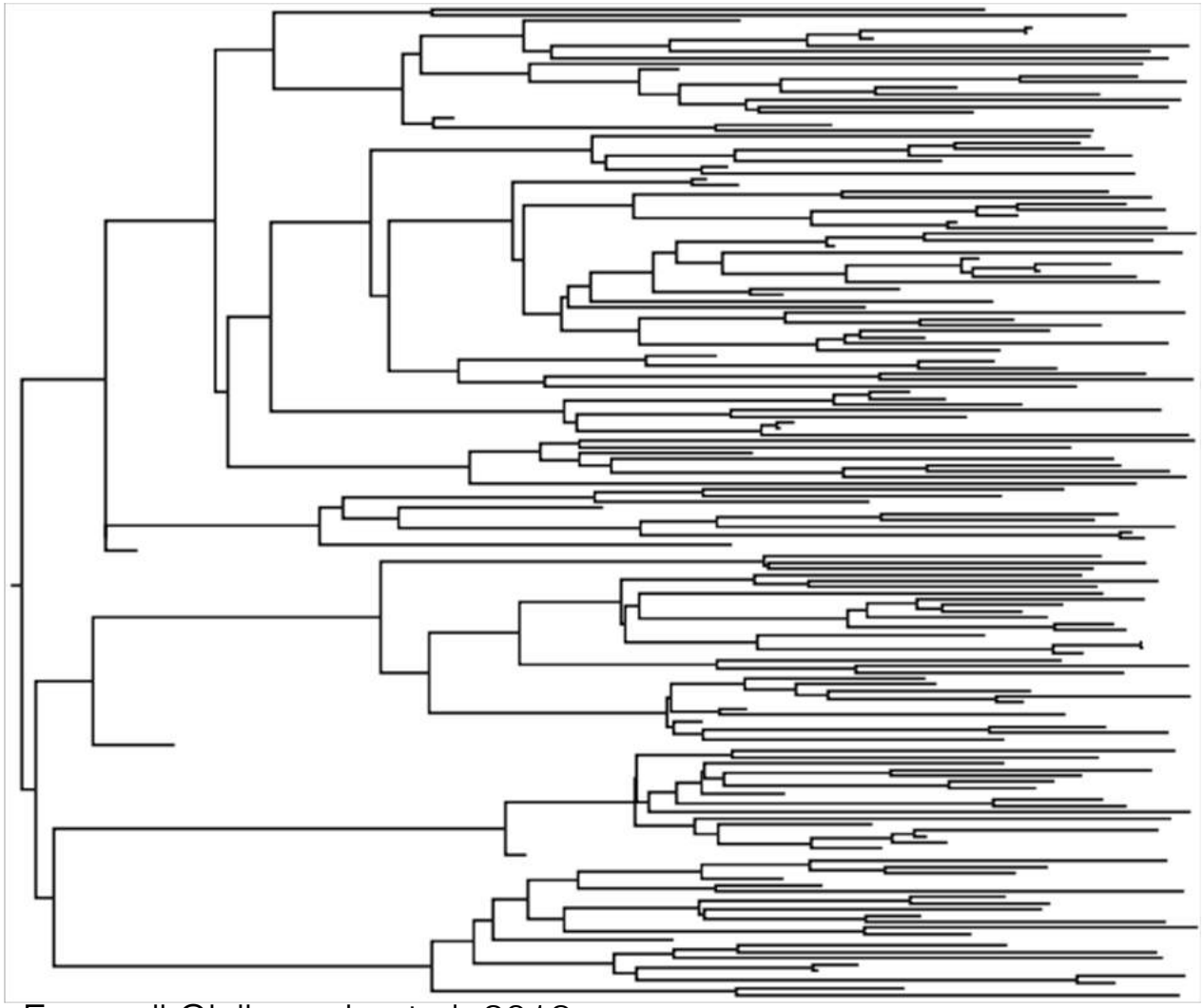
Tree prior and tree height (Zika virus outbreak)





From: Boskova et al. 2018

Tree prior and epi parameters

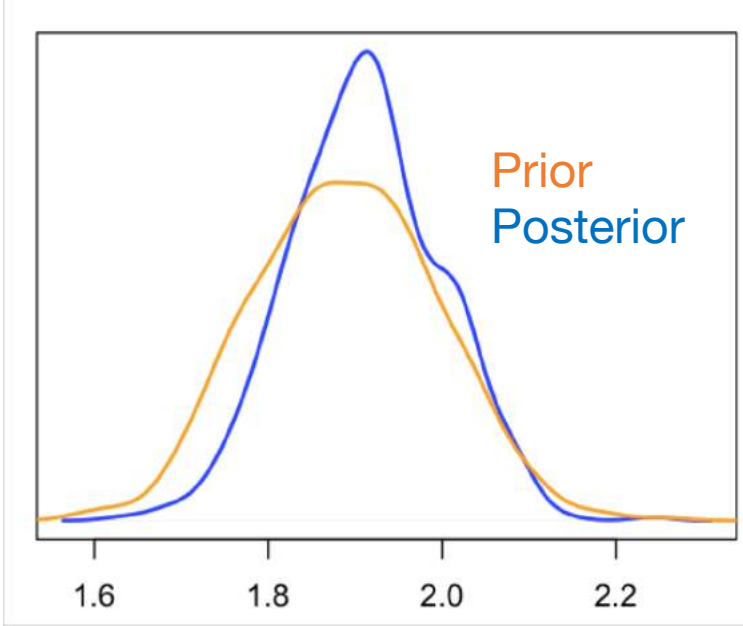


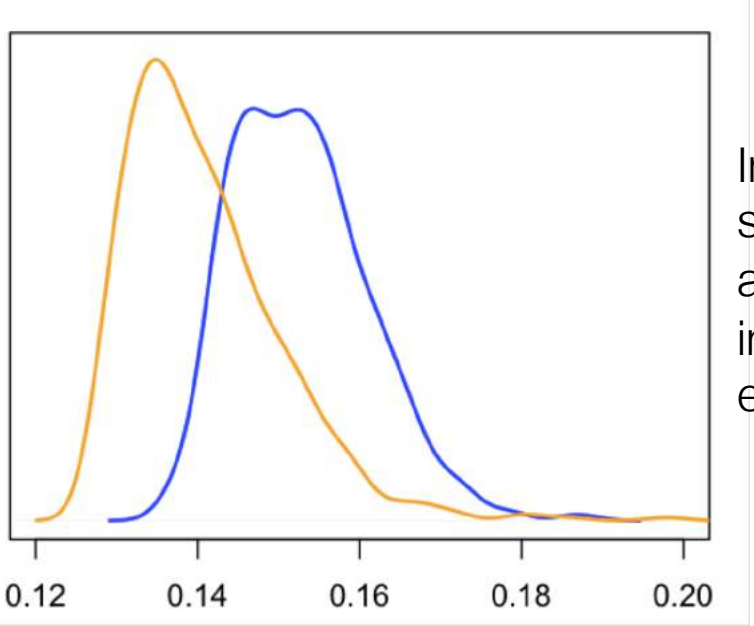
From: di Giallonardo et al. 2018

Data set: Human respiratory syncytial virus (RSVA)

- 163 samples collected over one month.
- Sampled in August (peak number of cases).
- Analyse using birth-death and coalescent exponential.

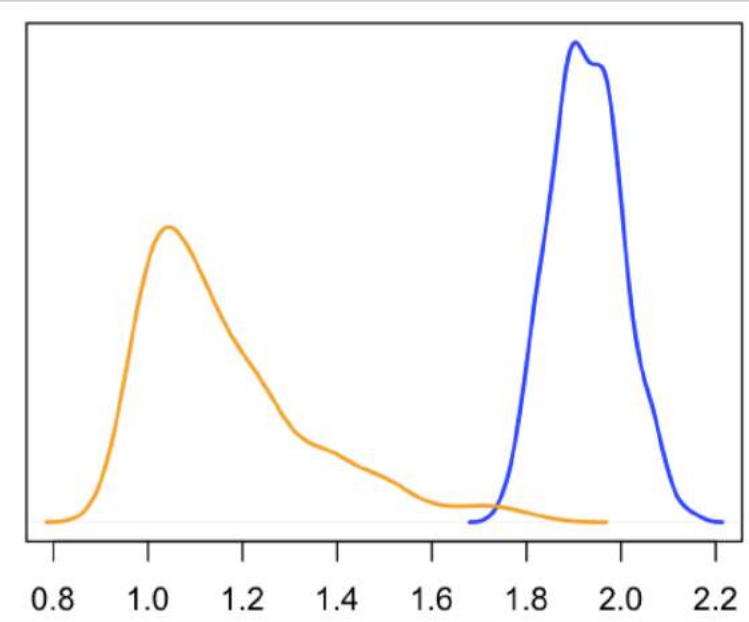
Birth-death



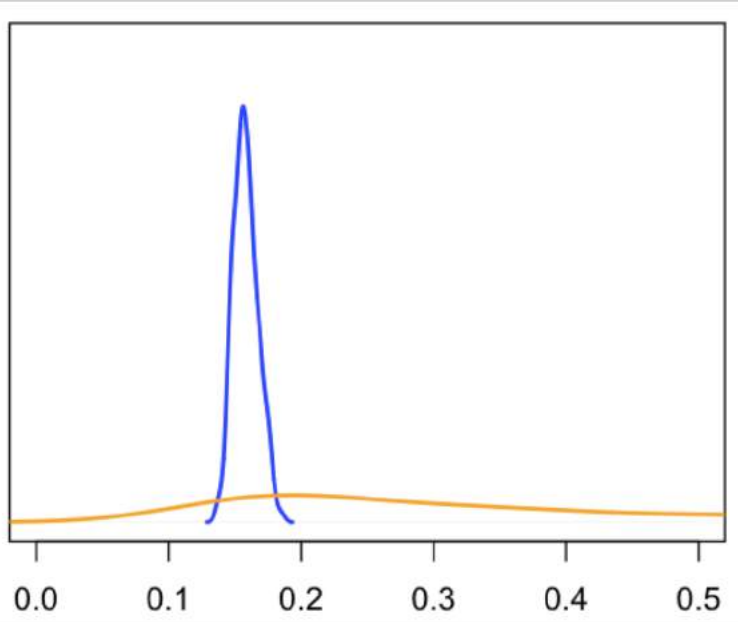


In the birth-death, sampling times alone are very informative about epi parameters.

Coalescent exponential

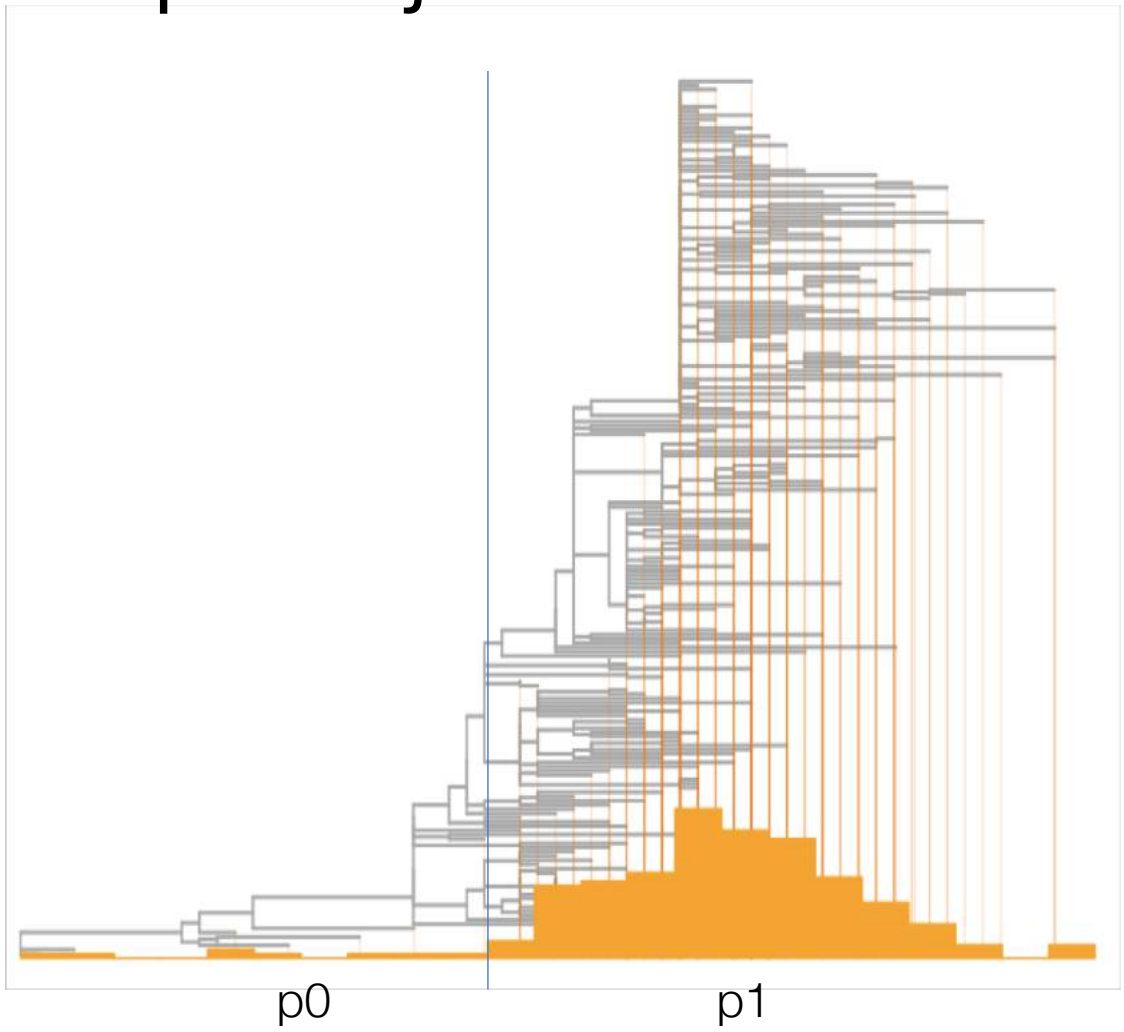


R_0



Root height (years)

Epi trajectories

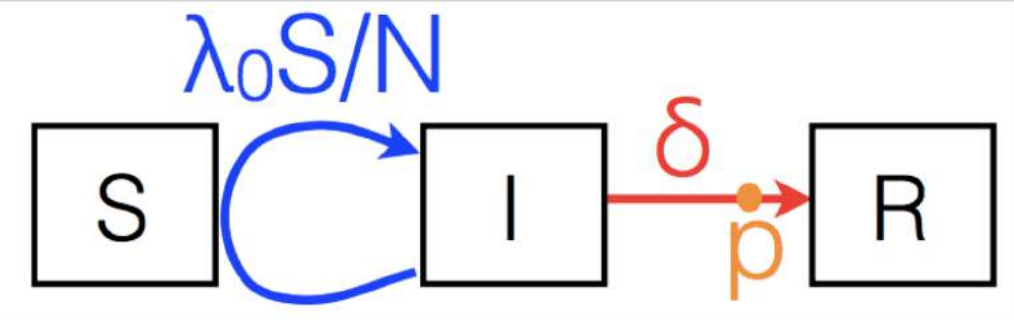


Data set: AH1N1 Influenza

- Samples collected in 2009 from February to August
- Estimate R_0 using three model:
  - Birth-death
  - Coalescent exp.
  - Birth-death SIR

Account for sampling heterogeneity using BDSIR (or BD skyline).

Susceptible-infected-recovered



$N=S+I+R$

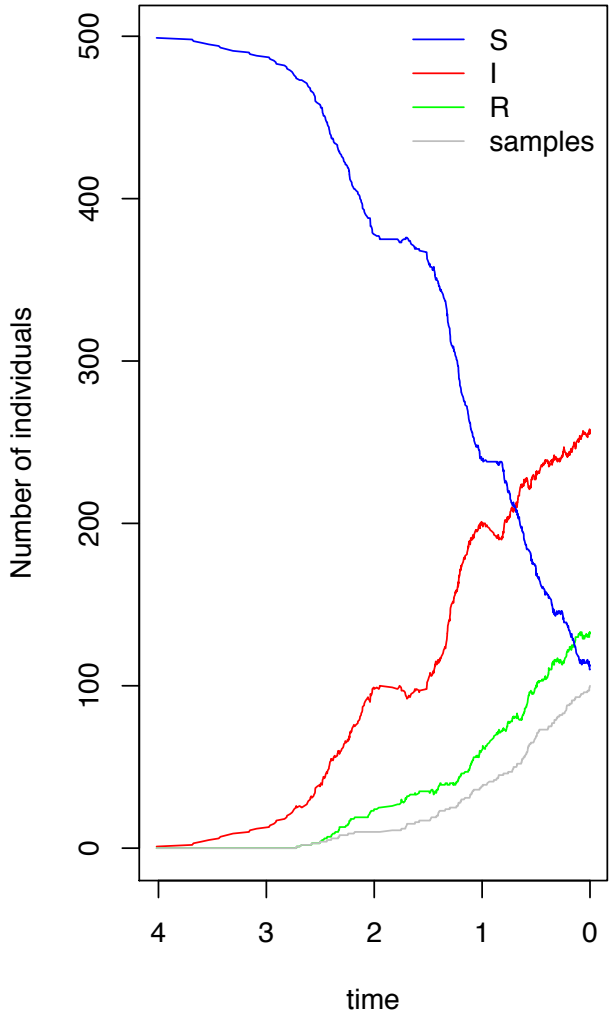
λ = transmission rate

δ = becoming non-infectious rate

ρ = sampling probability

Kuhnert et al. 2014
From: Stadler et al. 2014

Susceptible-infected-recovered



Kuhnert et al. 2014

Transmission rate (birth): β

Become uninfectious rate: γ

Susceptible pop. size: n_s

$R_e = \beta^* n_s / \gamma$

$R_0 = \beta^* n_s(0) / \gamma$

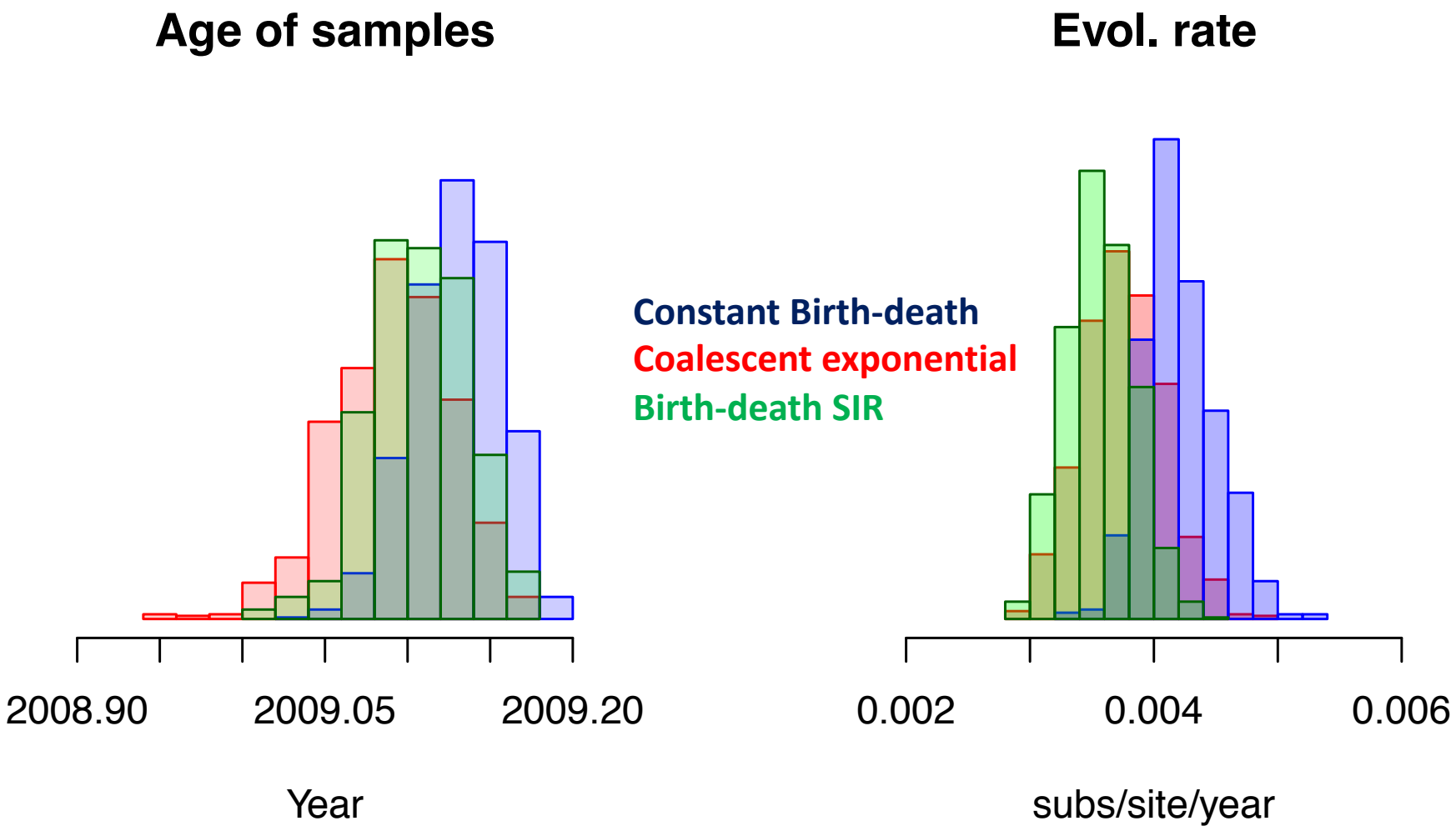
SIR and BD are similar when S → N

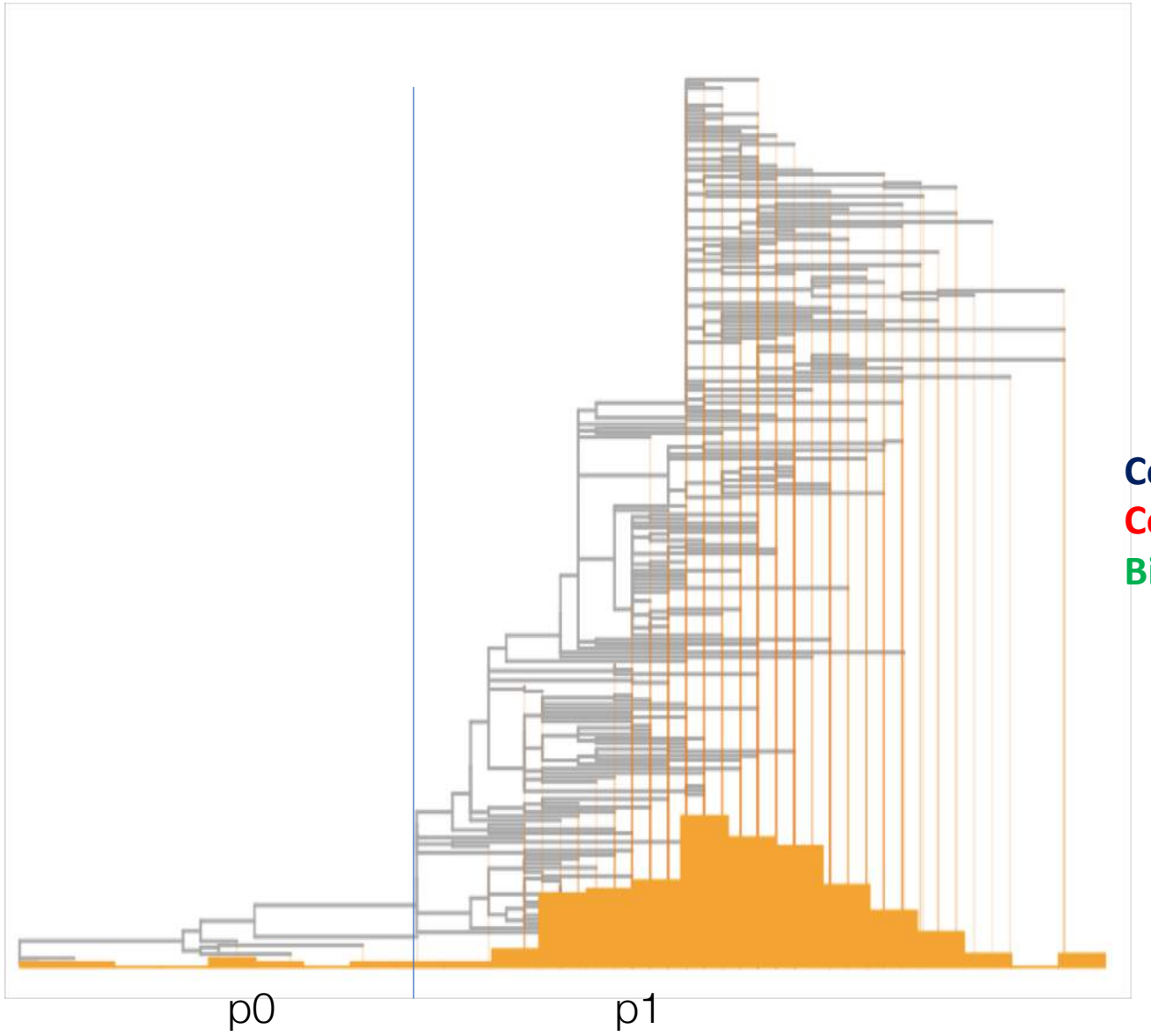
Comparing to birth-death notation:

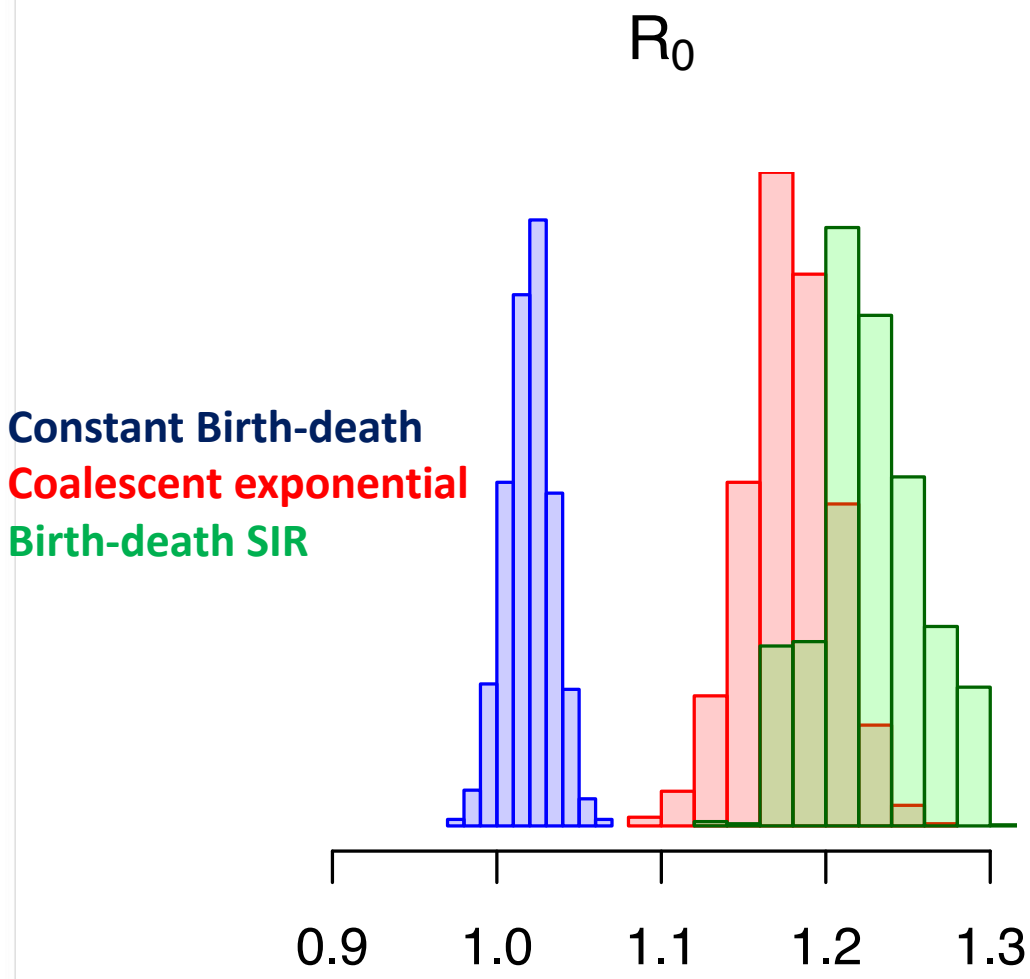
γ is equivalent to δ

$\beta^* n_s(0) = \lambda_0$

The birth-death SIR is an approximation of the SIR based on the birth-death skyline (also available; coalSIR, SIR).


Age of samples
Evol. rate
Constant Birth-death
Coalescent exponential
Birth-death SIR
2008.90
2009.05
2009.20
Year
0.002
0.004
0.006
subs/site/year

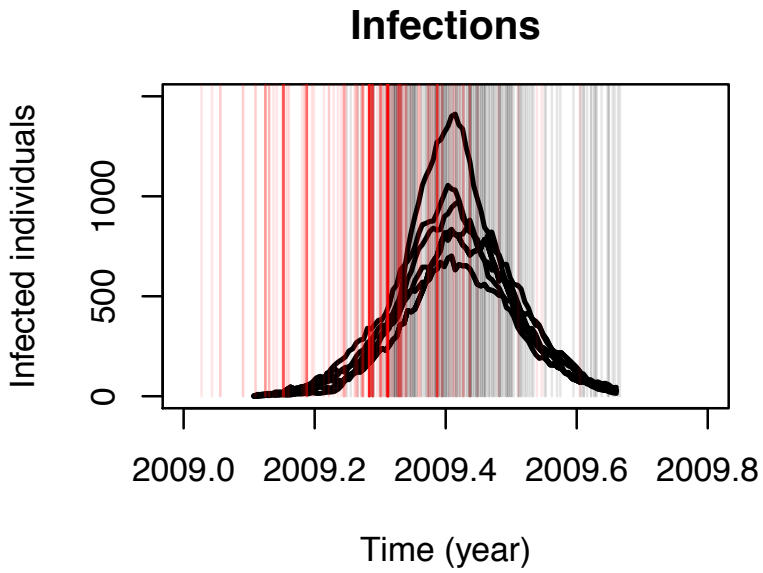


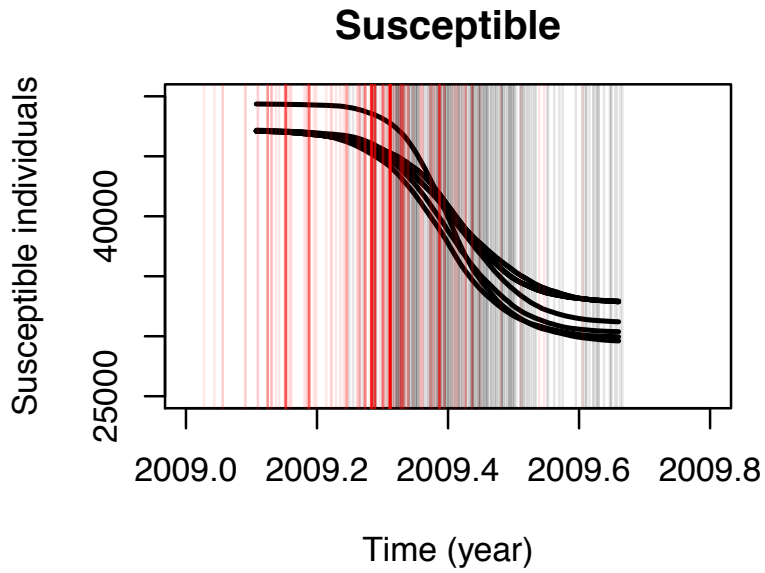


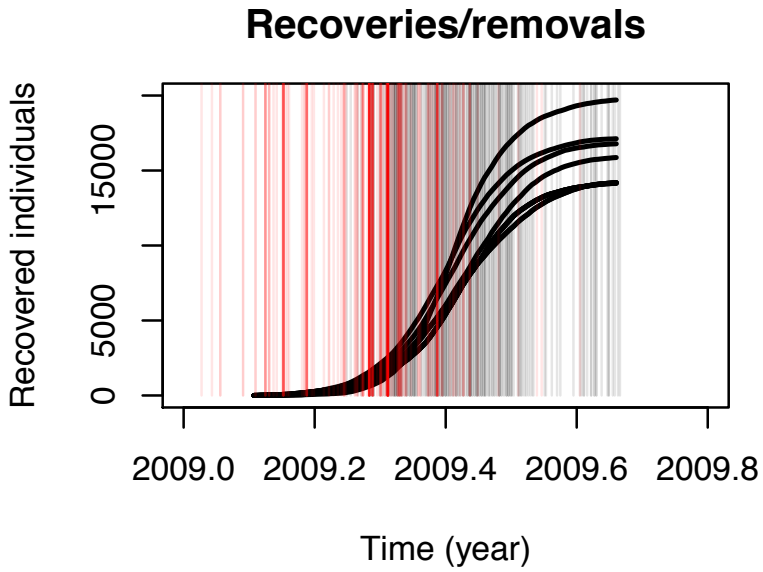
The **constant birth-death** assumes constant sampling over time!

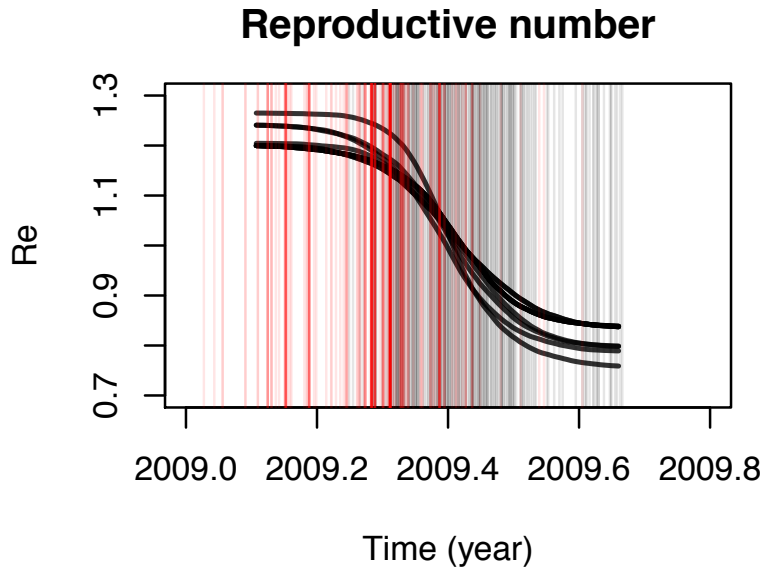
We can relax this assumption using the birth-death Skyline or the birth-death SIR.

SIR trajectories (BDSIR, CoalSIR, or SIR)







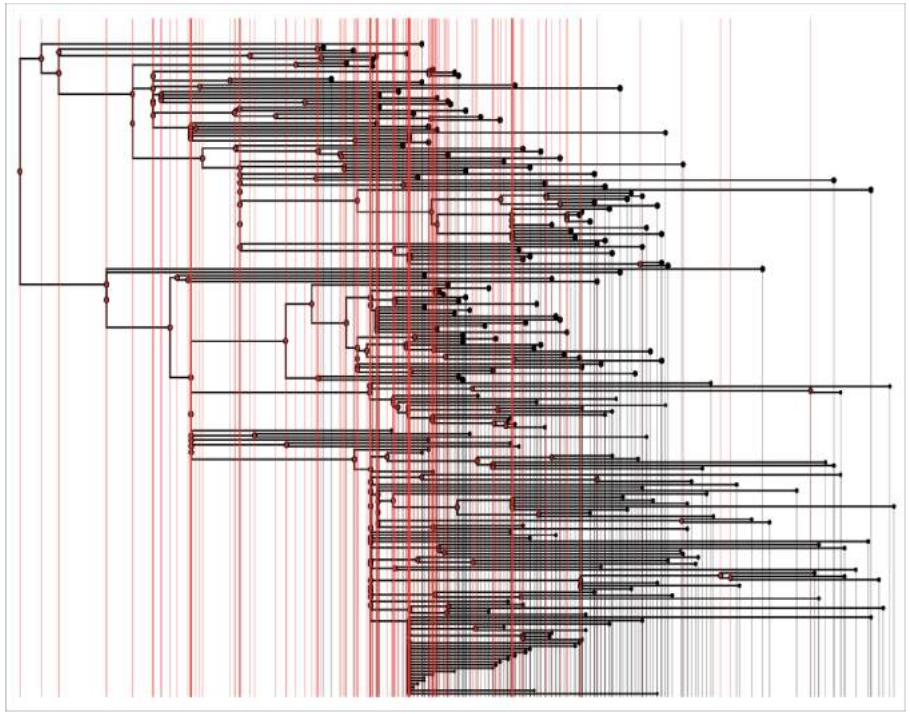


- Collected genome samples

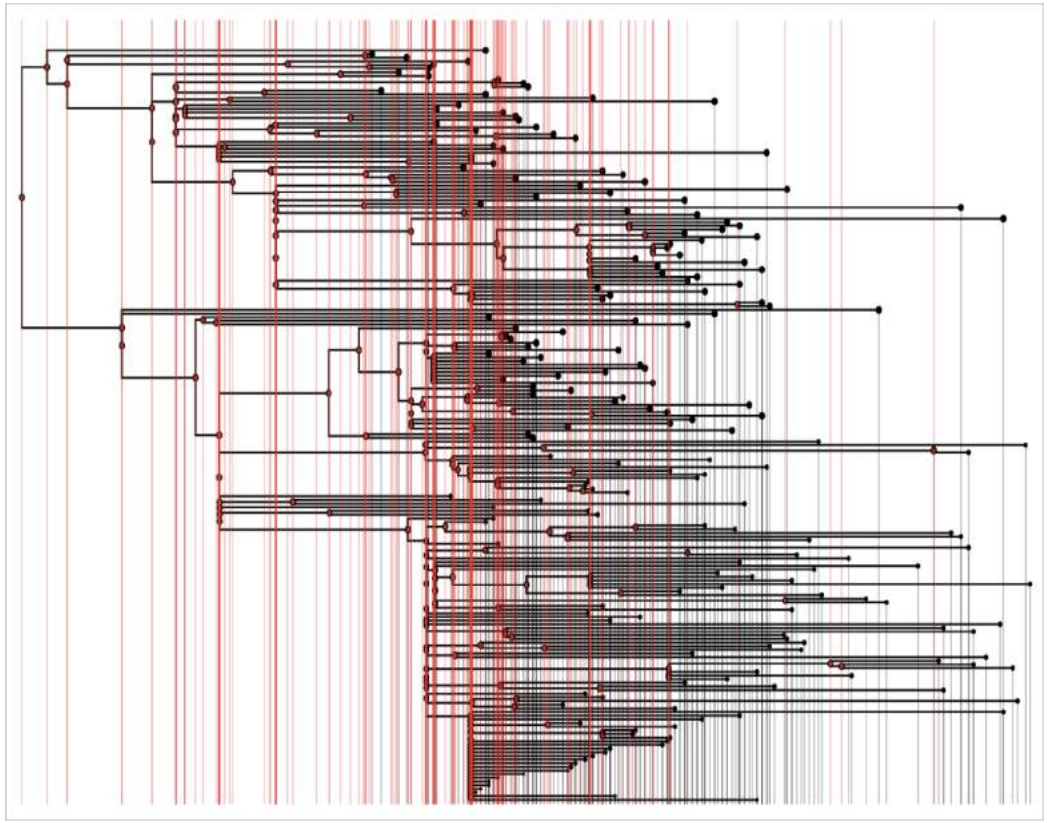
- Internal nodes (inferred transmission events)

Vaughan et al. 2017
Popinga et al. 2015
Kuhnert et al. 2014

- Tip-dates are a rich source of information
  - Calibration.
  - Information for some tree priors (e.g. birth-death).
- Verify temporal structure
  - Comparing prior and posteriors.
  - Tip-dates can be specified using probability distributions (e.g. ancient DNA).



- Epidemiological inferences
  - Internal node information.
  - Branching as an epidemiological process
  - Check phylodynamic model assumptions
    - E.g. via model adequacy; phyloseminar.org on model adequacy of infectious disease phylodynamics and Duchene et al. 2018.

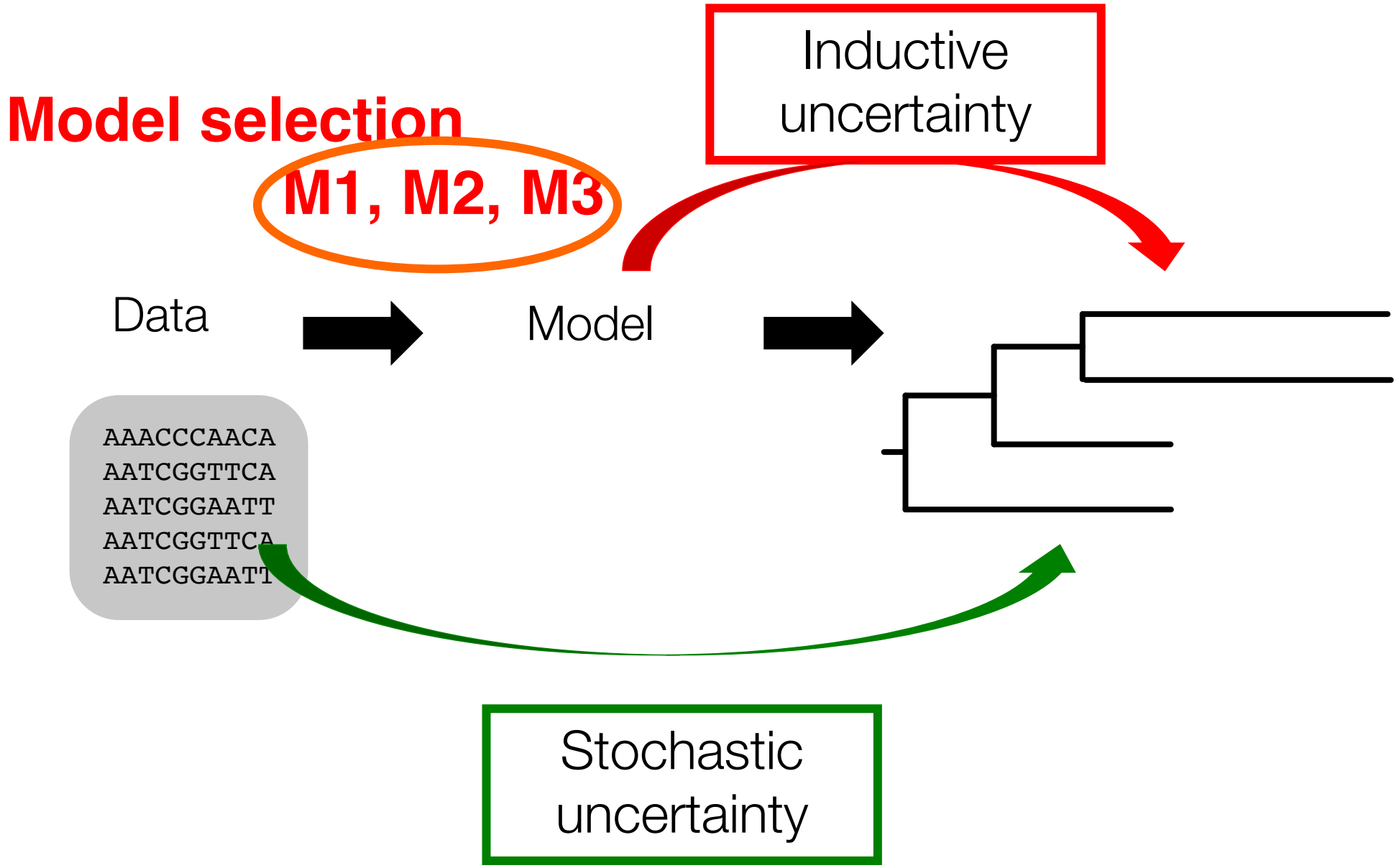


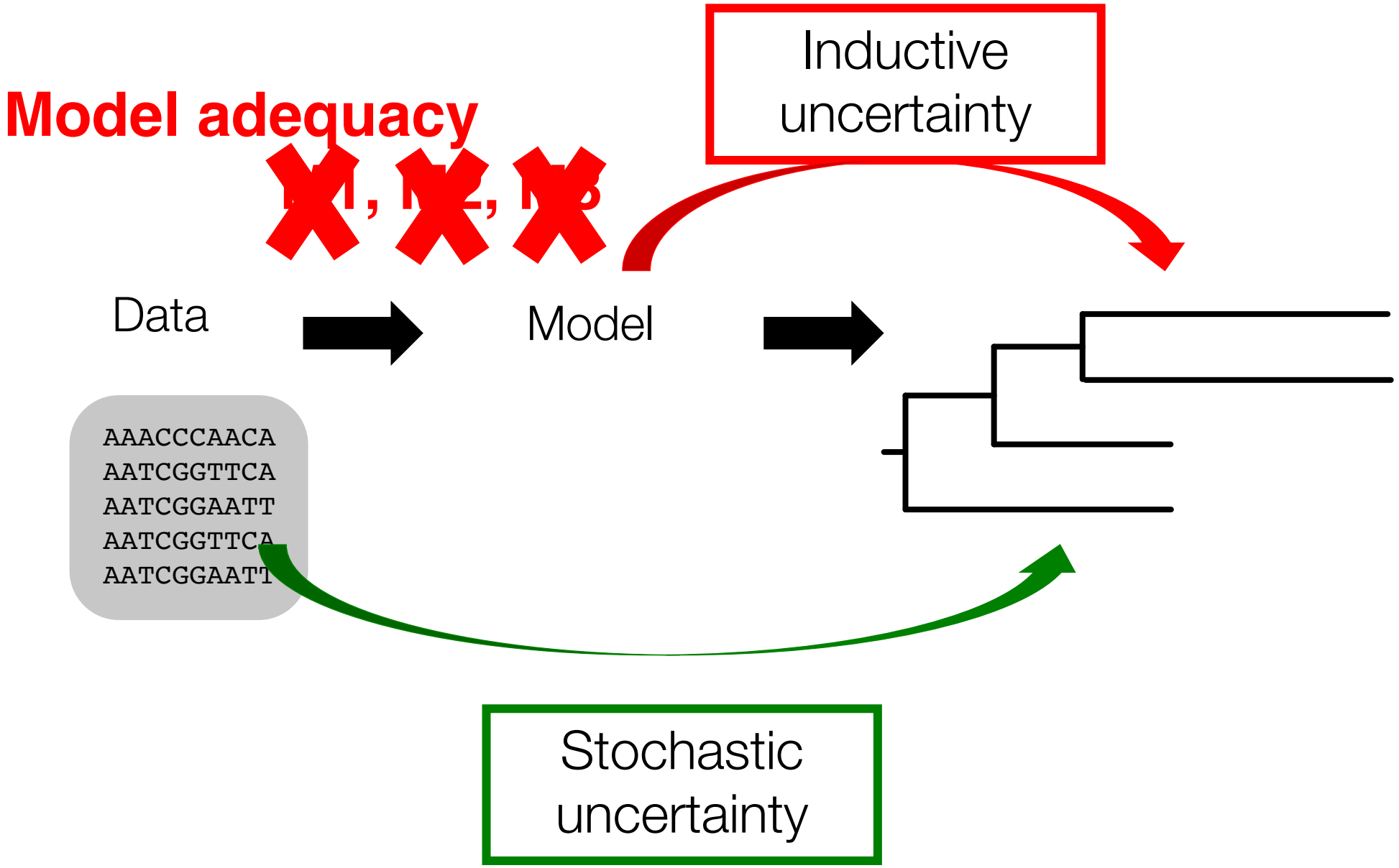
- In the future…
  - Hospital notification data.
  - Drug resistance profiles.
  - Virulence.

Some useful references

- Boskova, V., Stadler, T., & Magnus, C. (2018). The influence of phylodynamic model specifications on parameter estimates of the zika virus epidemic. Virus evolution, 4(1), vex044.
- Du plessis, l., & Stadler, T. (2015). Getting to the root of epidemic spread with phylodynamic analysis of genomic data. Trends in microbiology, 23(7), 383-386.
- Ho, s. Y., & Duchêne, S. (2014). Molecular-clock methods for estimating evolutionary rates and timescales. Molecular ecology, 23(24), 5947-5965.
- Duchêne, s., Duchêne, D., Holmes, E. C., & Ho, S. Y. (2015). The performance of the date-randomization test in phylogenetic analyses of time-structured virus data. Molecular biology and evolution, 32(7), 1895-1906.

Model adequacy and phylodynamics


Model selection
M1, M2, M3
Inductive uncertainty
Data
Model
AAACCCAACA
AATCGGTTCA
AATCGGAATT
AATCGGTTCA
AATCGGAATT
Stochastic uncertainty


Model adequacy
M1, M2, M3
Inductive uncertainty
Data
Model
AAACCCAACA
AATCGGTTCA
AATCGGAATT
AATCGGTTCA
AATCGGAATT
Stochastic uncertainty

Model selection

Select a pool of models

Rank models according to statistical fit

Select that with the highest fit

Model adequacy

Consider a model(s)

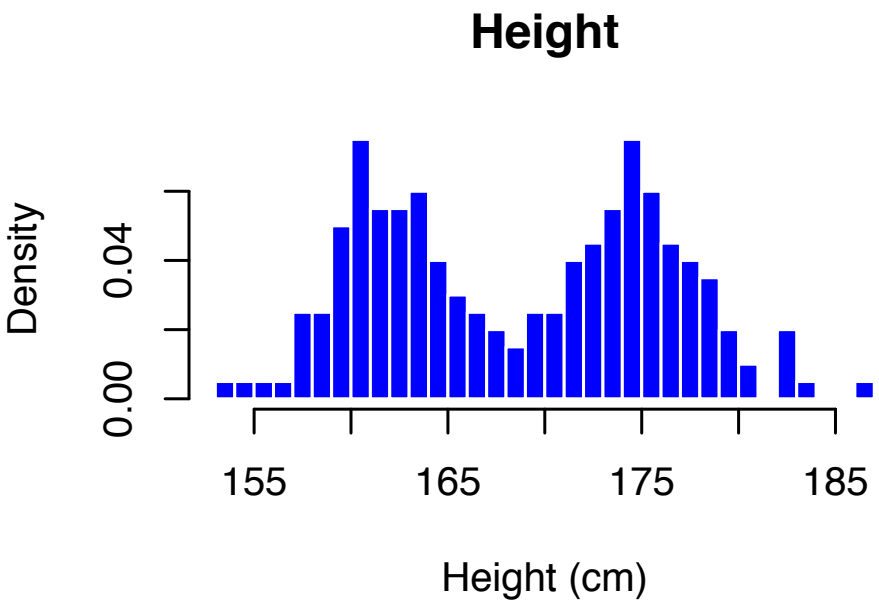
Treat models as hypotheses

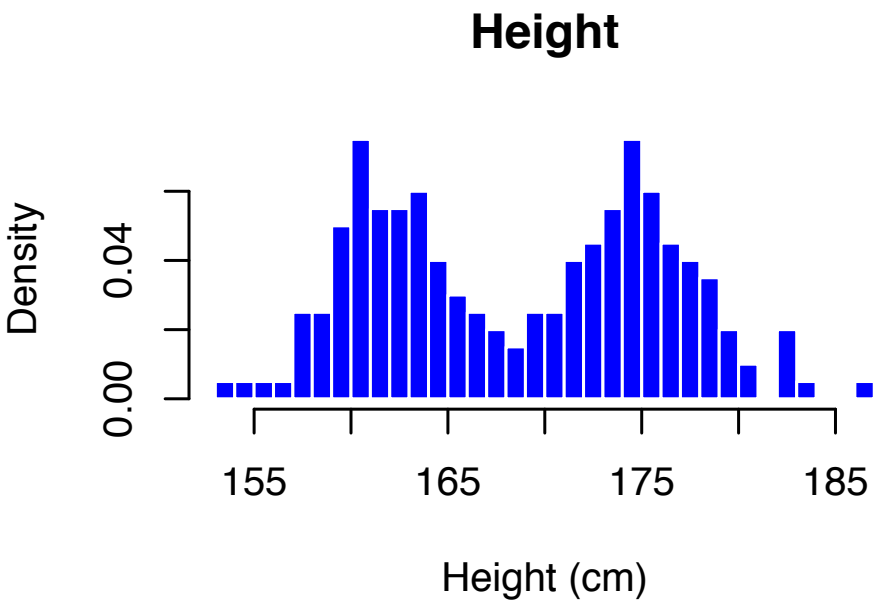
Conduct hypothesis testing to assess the limits of the model's applicability

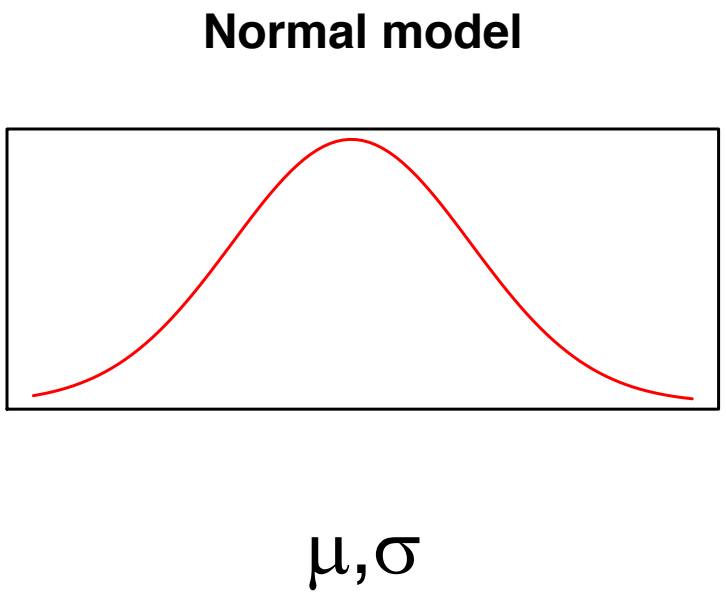
The goal of model adequacy

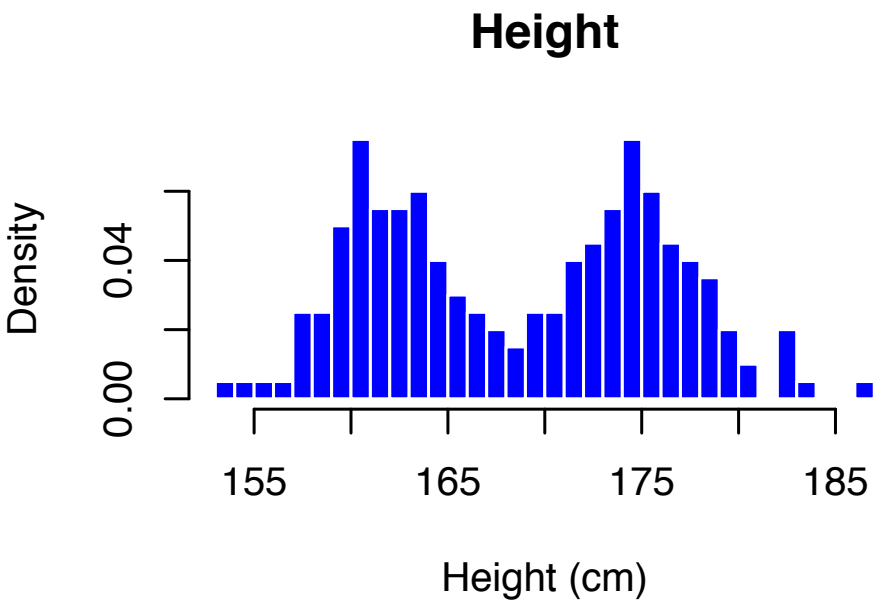
- *The relevant goal is not to answer the question, 'Do the data come from the assumed model?' (to which the answer is almost always no), but to* ***quantify the discrepancies between data and model****, and assess whether they could have arisen by chance, under the model's own assumptions.*

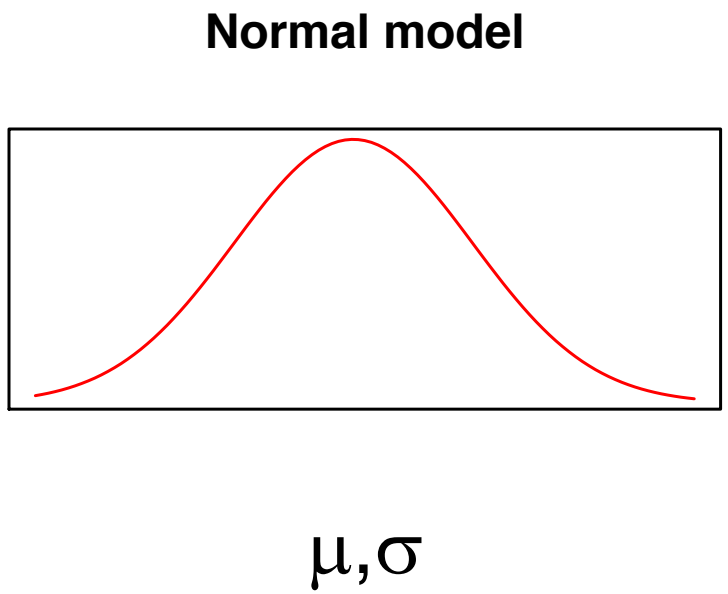
Gelman et al. 2013
Chapter 6 of Bayesian Data Analysis, 3rd Edition

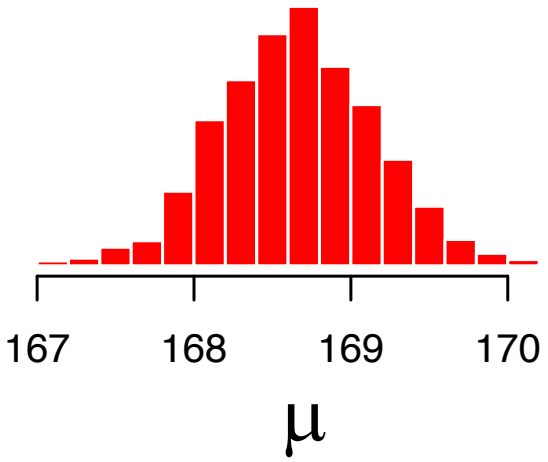

Height
Density
0.00
0.04
155
165
175
185
Height (cm)

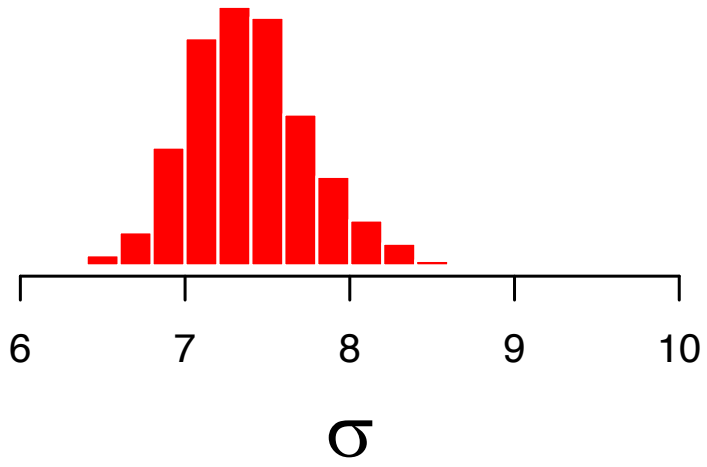

Height
Density
0.00
0.04
155
165
175
185
Height (cm)

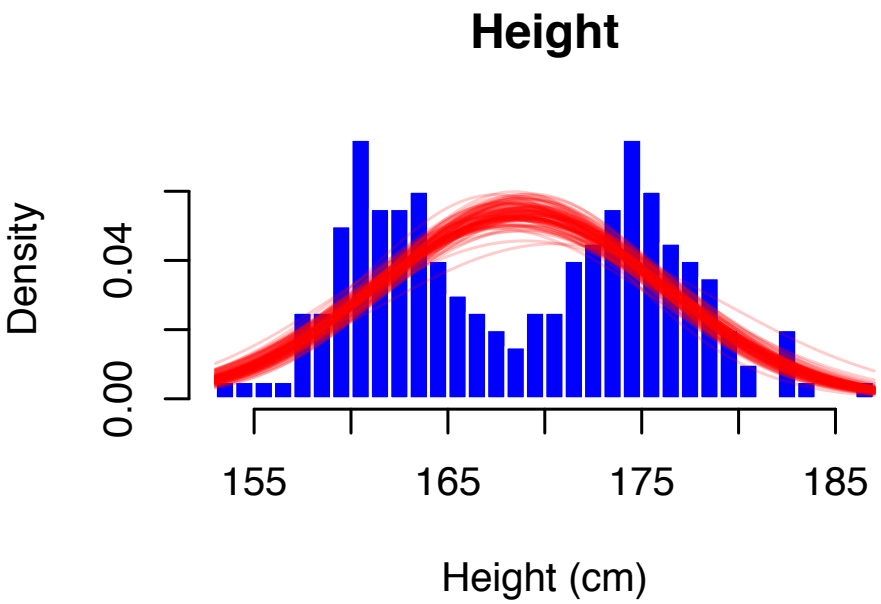

Normal model
μ,σ

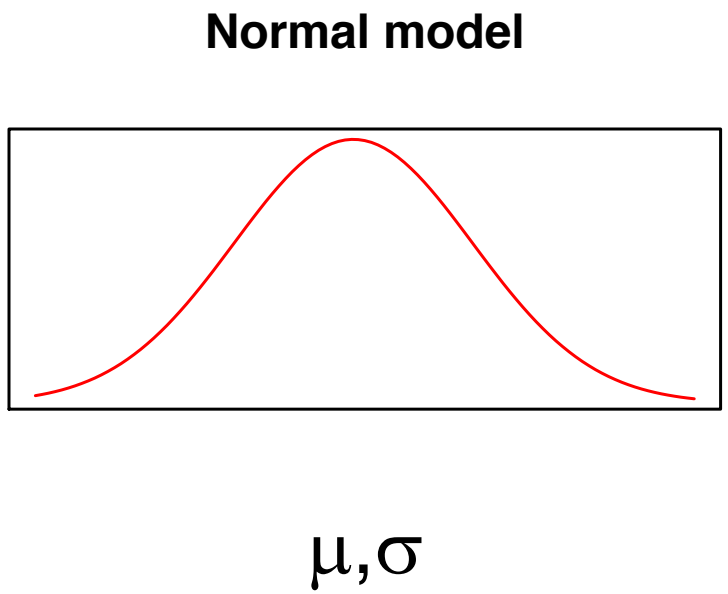

Height
Density
0.00
0.04
155
165
175
185
Height (cm)

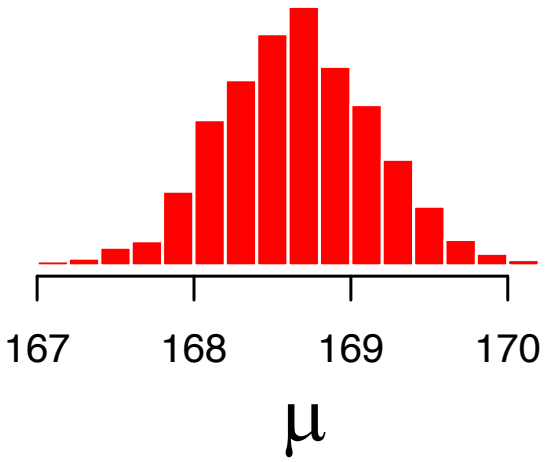

Normal model
μ,σ

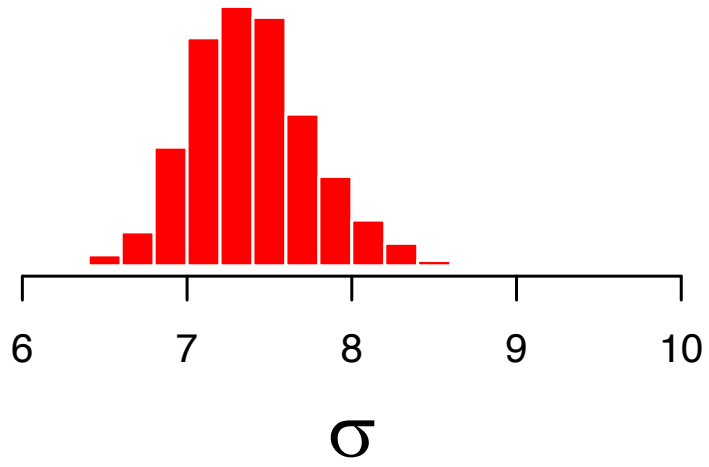

167
168
169
170
μ


6
7
8
9
10
σ

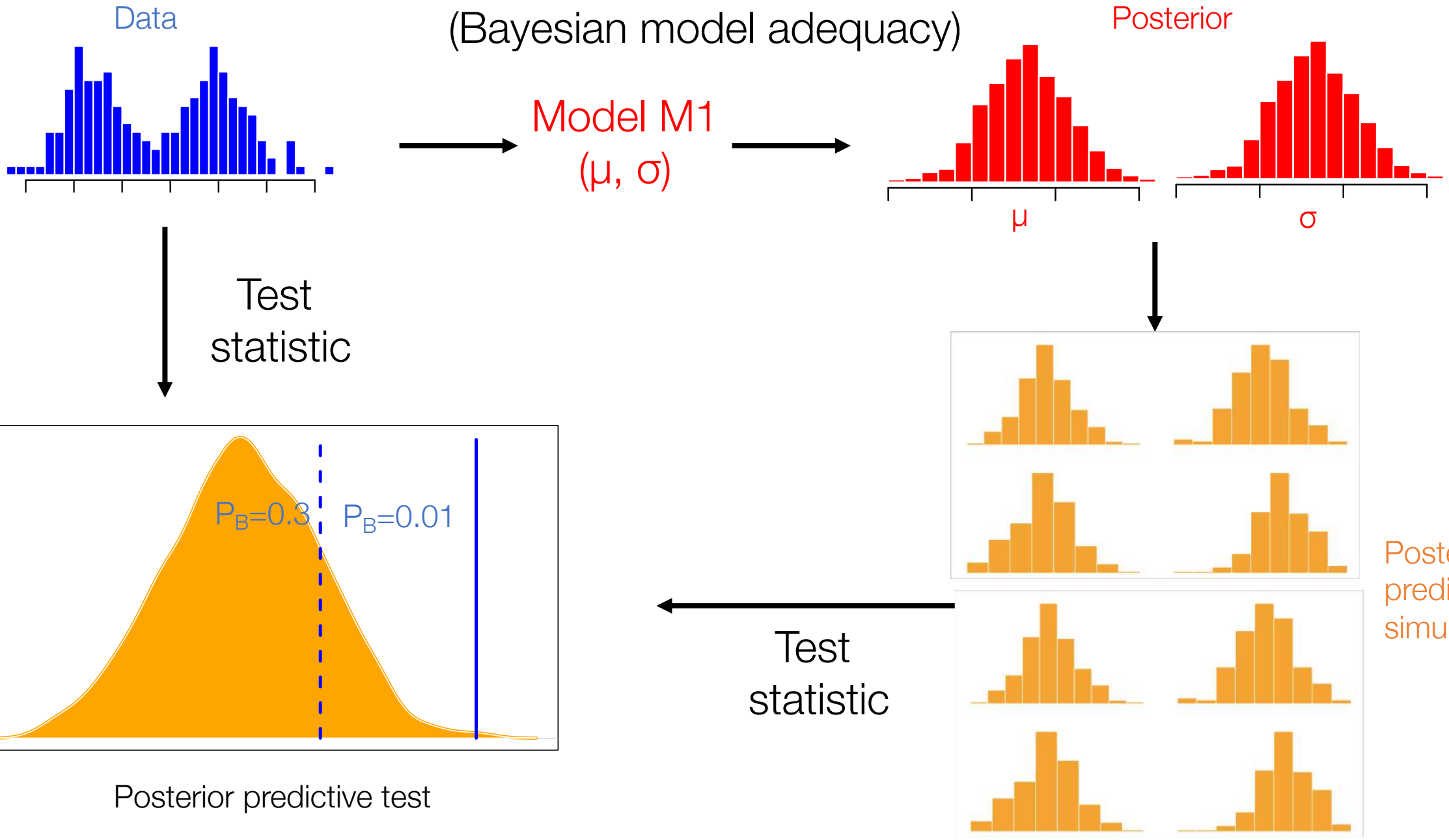

Height
Density
0.00
0.04
155
165
175
185
Height (cm)


Normal model
μ,σ

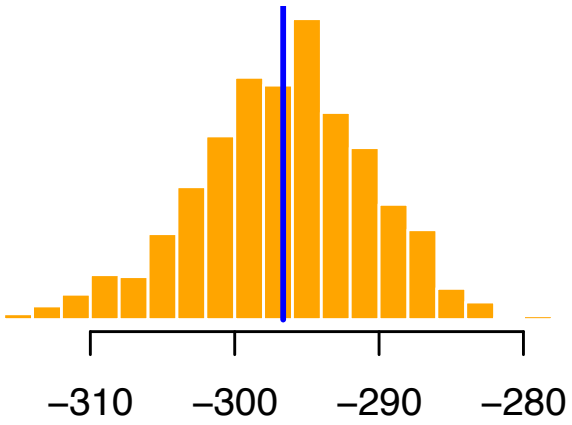

167
168
169
170
μ


6
7
8
9
10
σ

Posterior predictive model checking (Bayesian model adequacy)

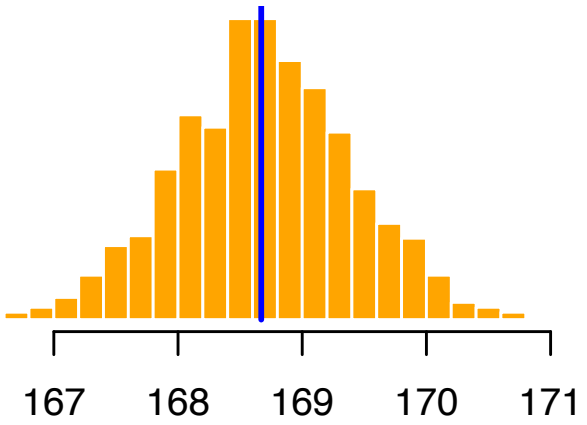


Mean likelihood



$P_B = 0.481$

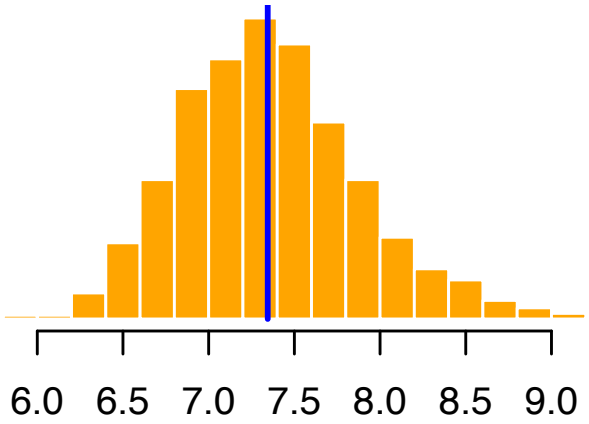
Mean



$P_B = 0.494$

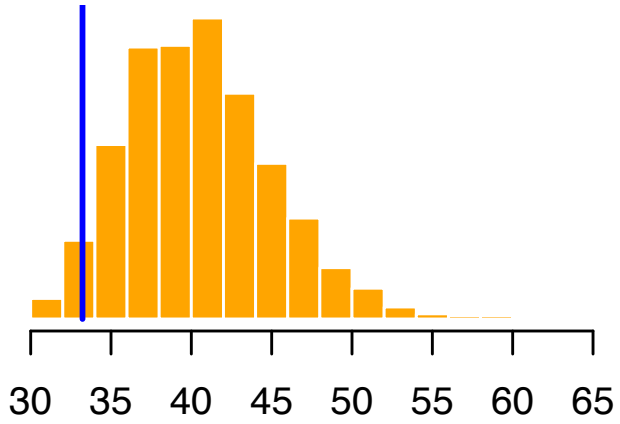
Gelman et al. 2013
Chapter 6 of Bayesian Data Analysis, 3rd Edition.

Sd



$P_B = 0.522$

Range



$P_B = 0.019$

"Sufficient statistics are bad test statistics"

On multiple comparisons…

- We can account for multiple comparisons (e.g. Bonferroni correction or multidimensional p-value).

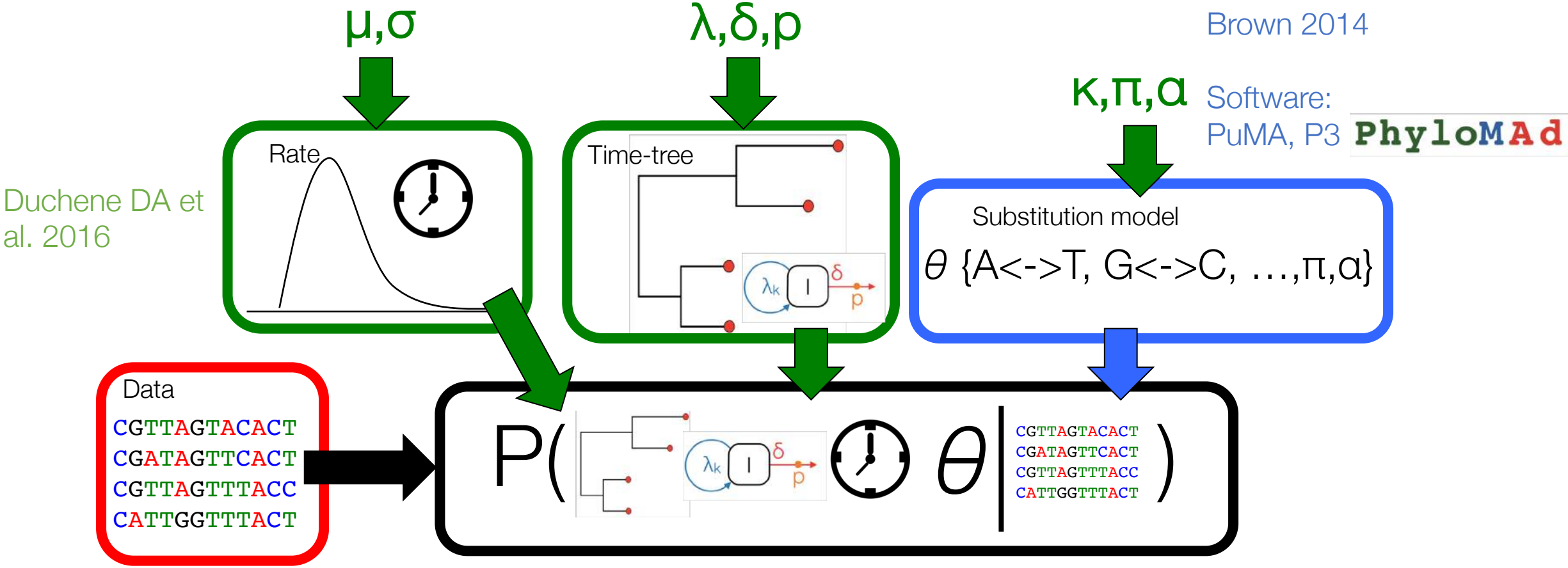
- However, we do not make this adjustments because the goal of model adequacy is to assess how the model predicts particular aspects of the data.

Tree prior/ branching model /phylodynamic model
Duchene S et al. 2018, Revell et al. 2005,
Drummond and Suchard 2008
Höhna 2015, Pennell et al. 2015

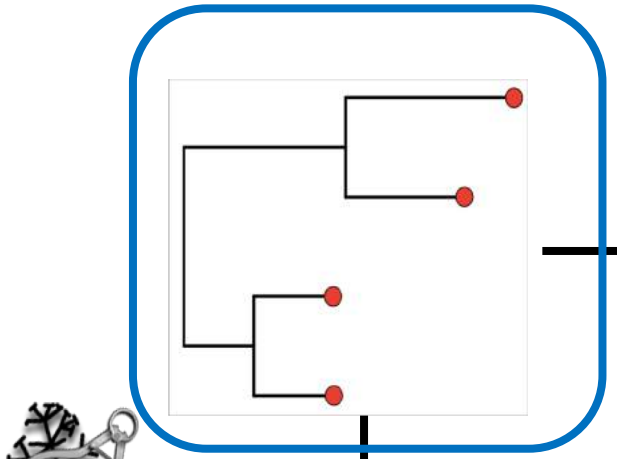
E.g.
Reeves 1992
Goldman 1993
Bollback 2002
Brown 2014

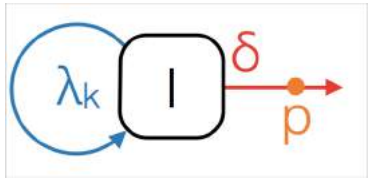
Software:
PuMA, P3 PhyloMAd

Duchene DA et al. 2016

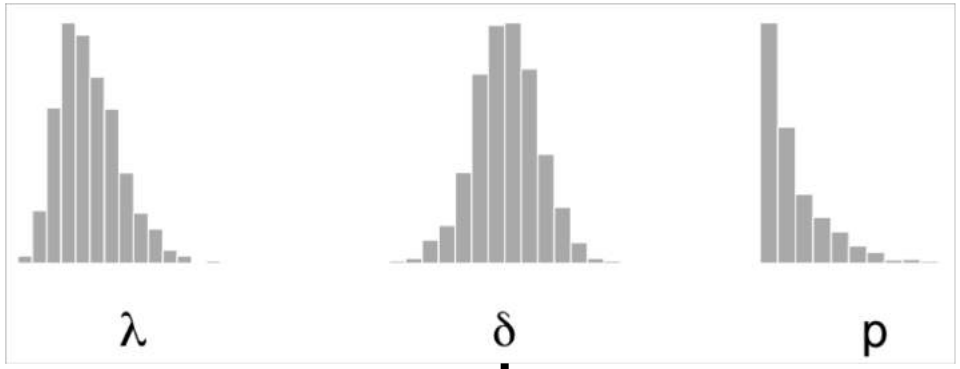


$$P(\eta \mid \Psi) \propto P(\Psi \mid \eta)\, P(\eta)$$



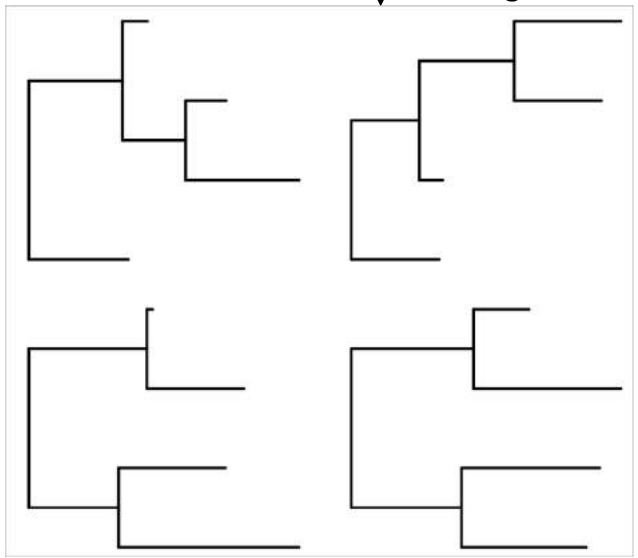


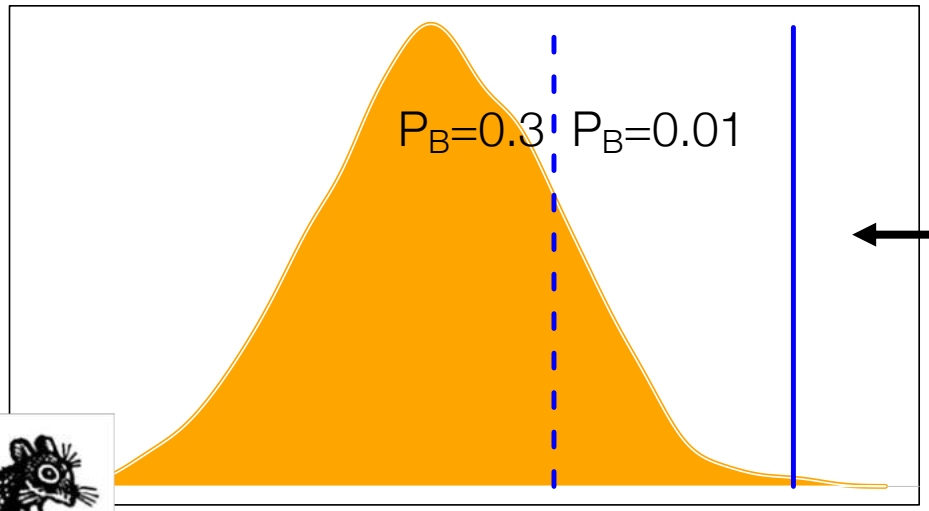
$\eta : \{\lambda, \delta, p\}$



MASTER

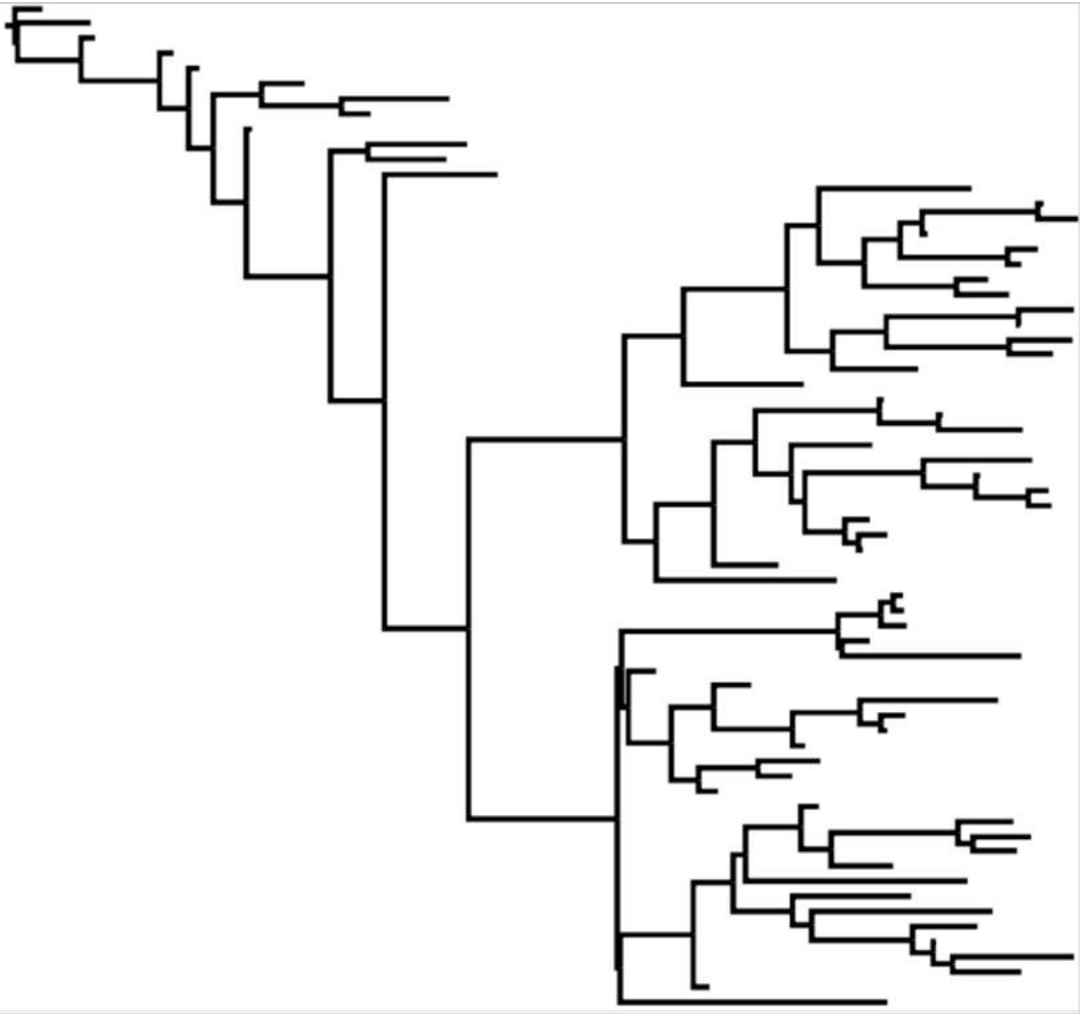
Vaughan et al. 2013



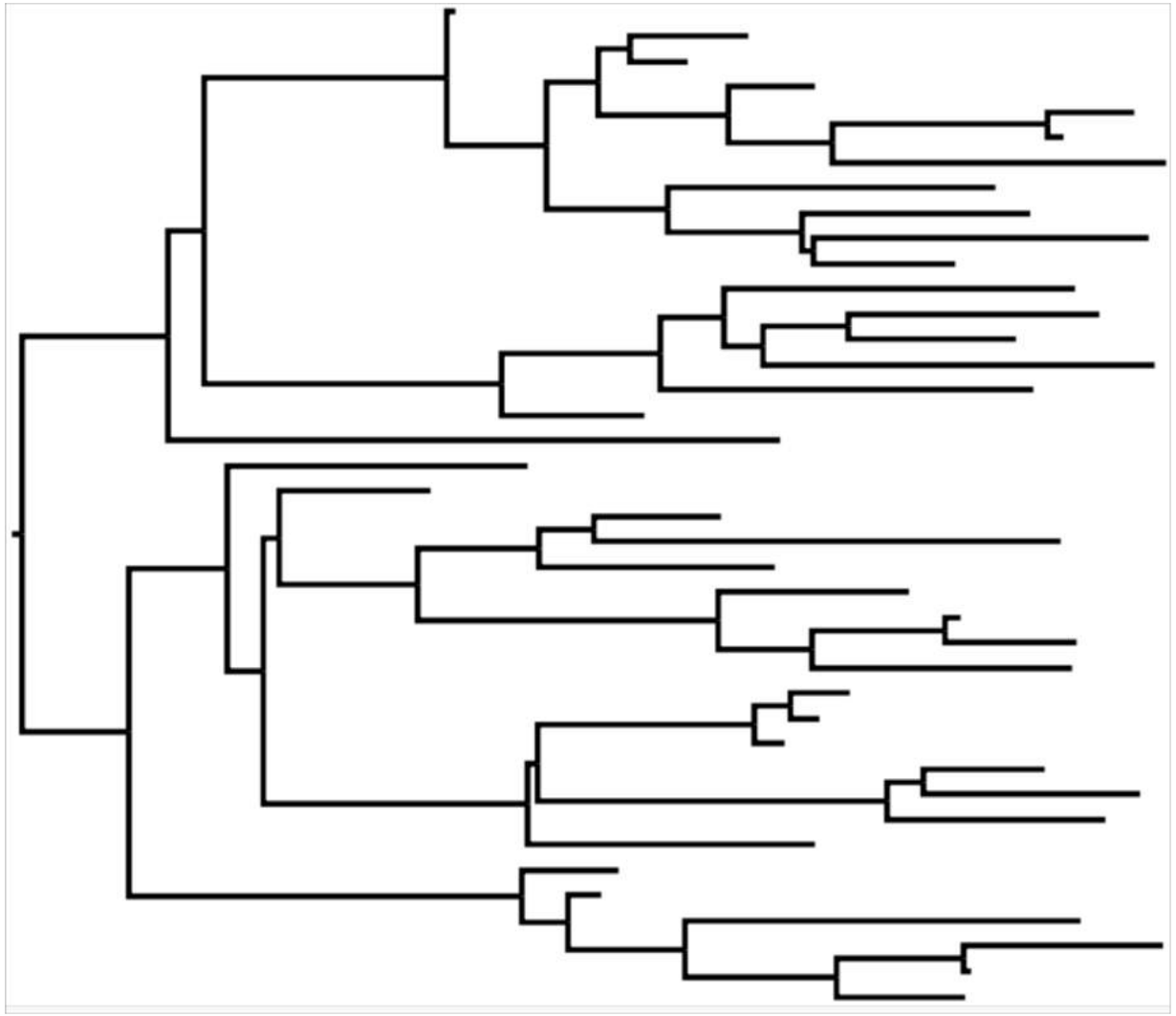


BEAST 2 package TreeModelAdequacy: Duchene et al. 2018

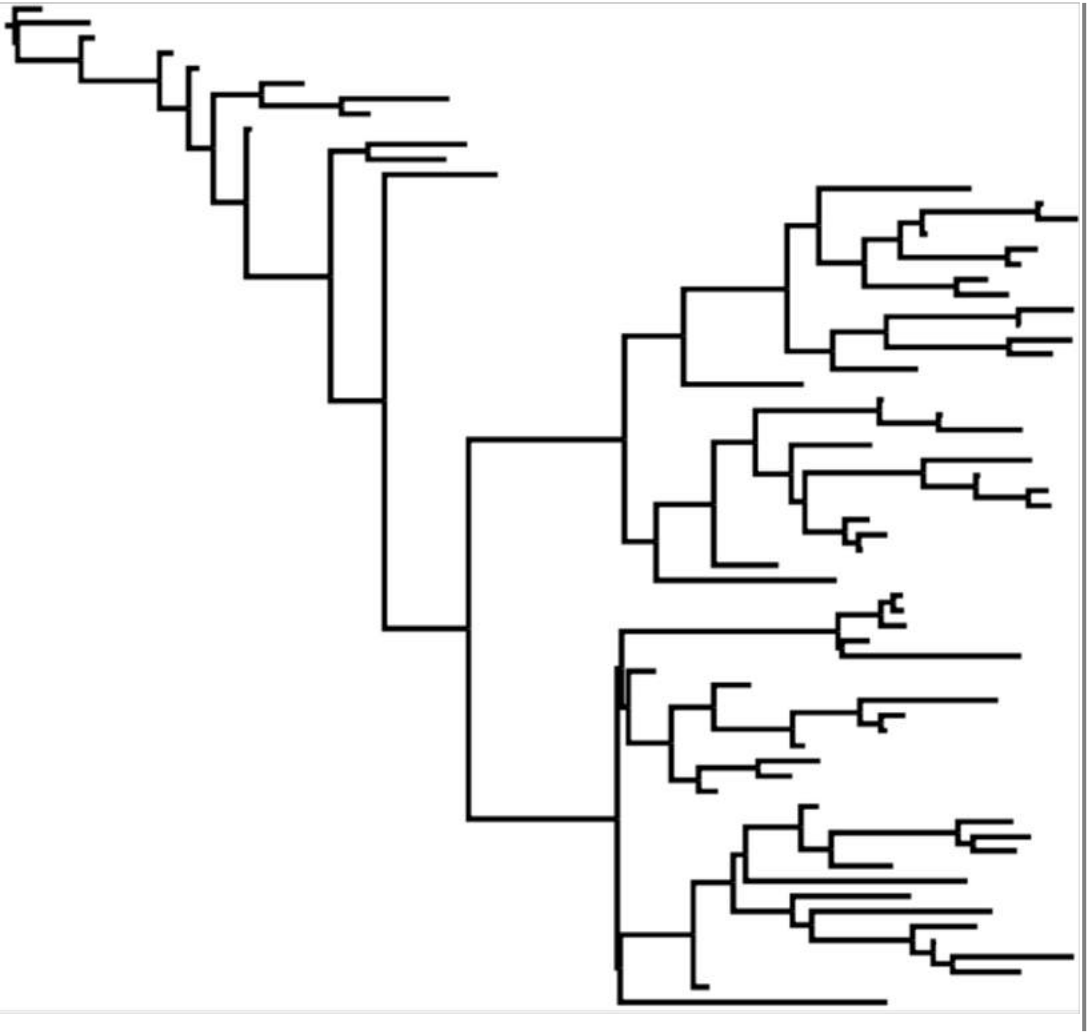
Birth-death constant sampling



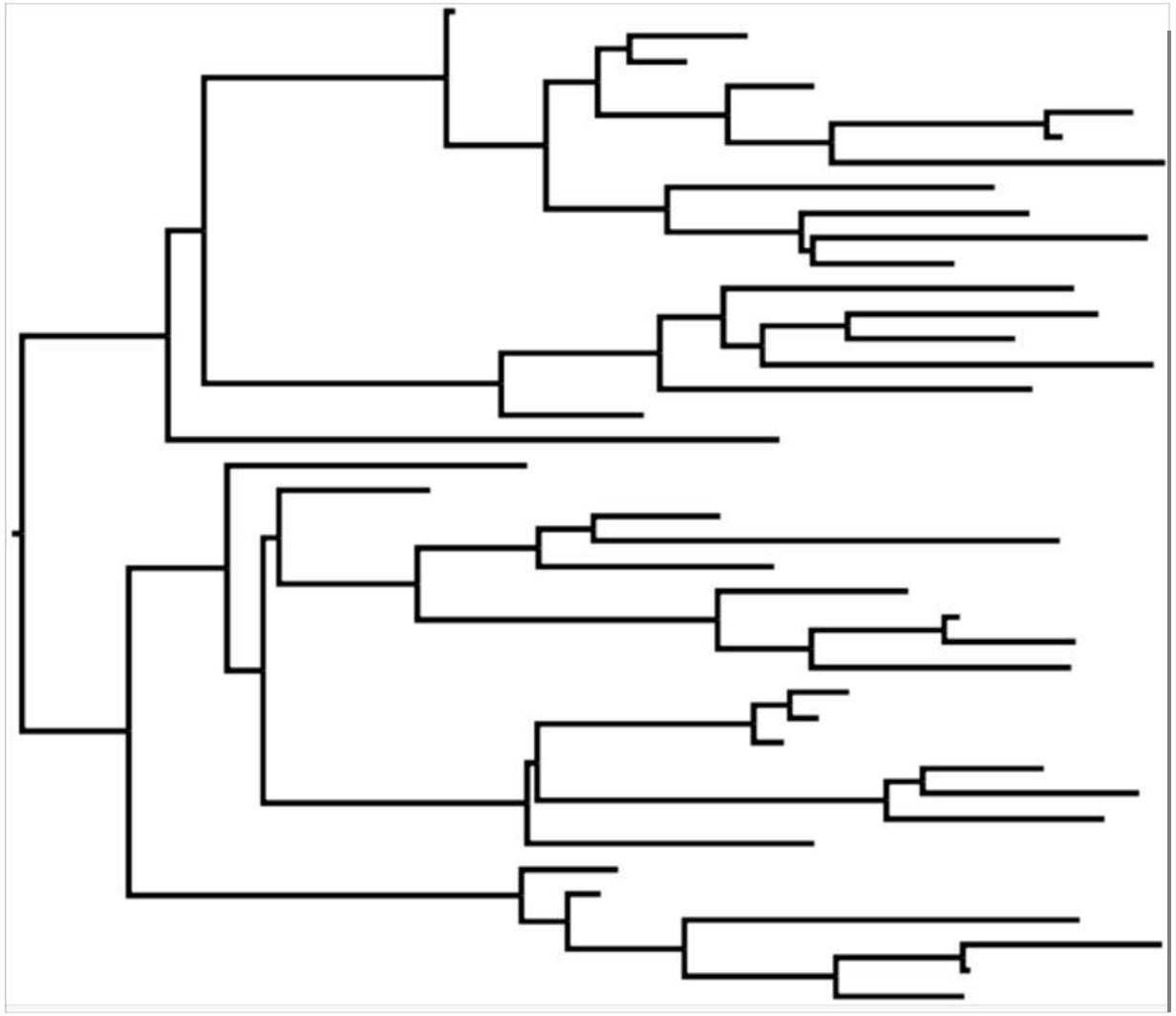
Birth-death skyline- variable sampling



Birth-death constant sampling

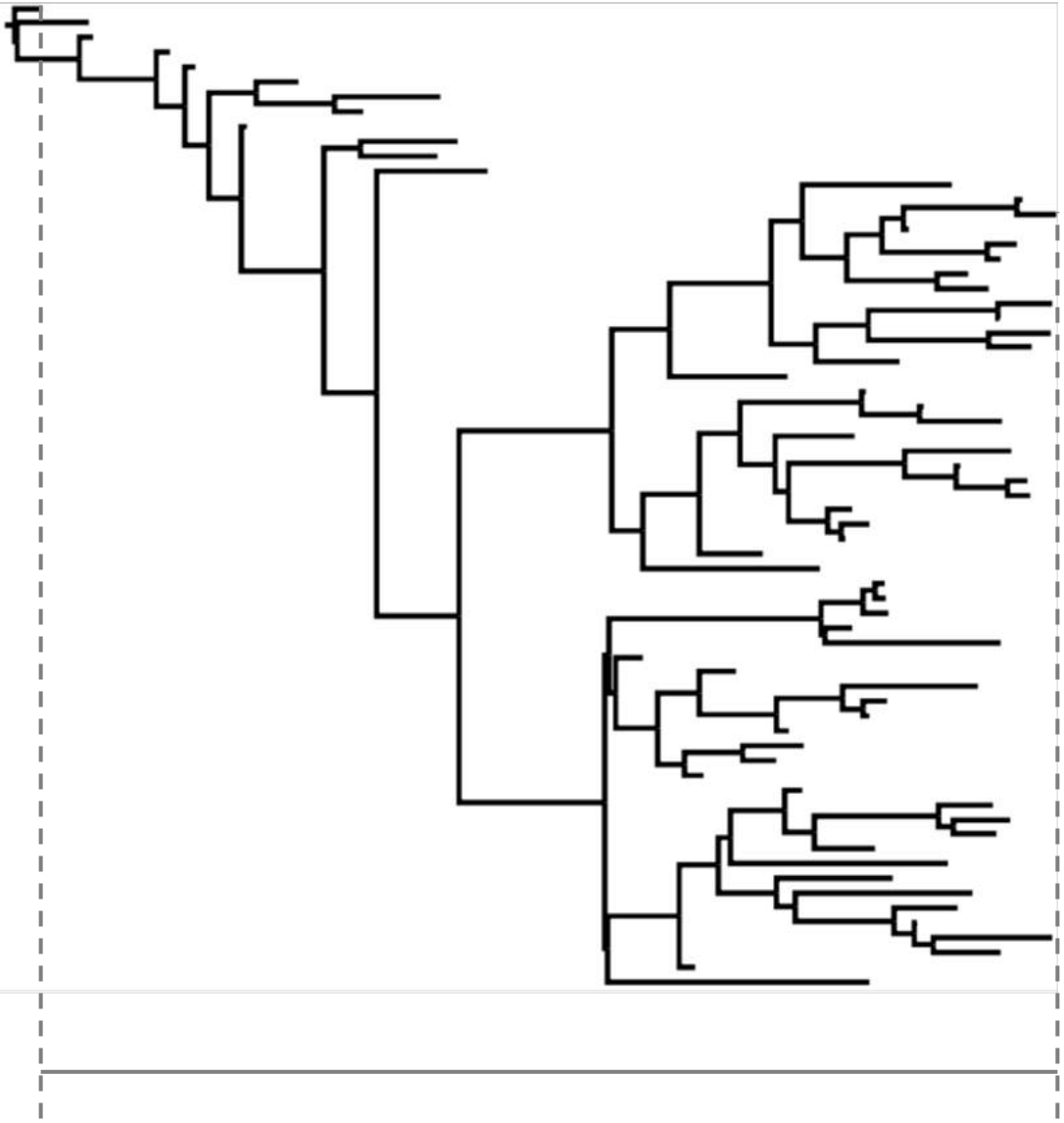


Birth-death skyline- variable sampling



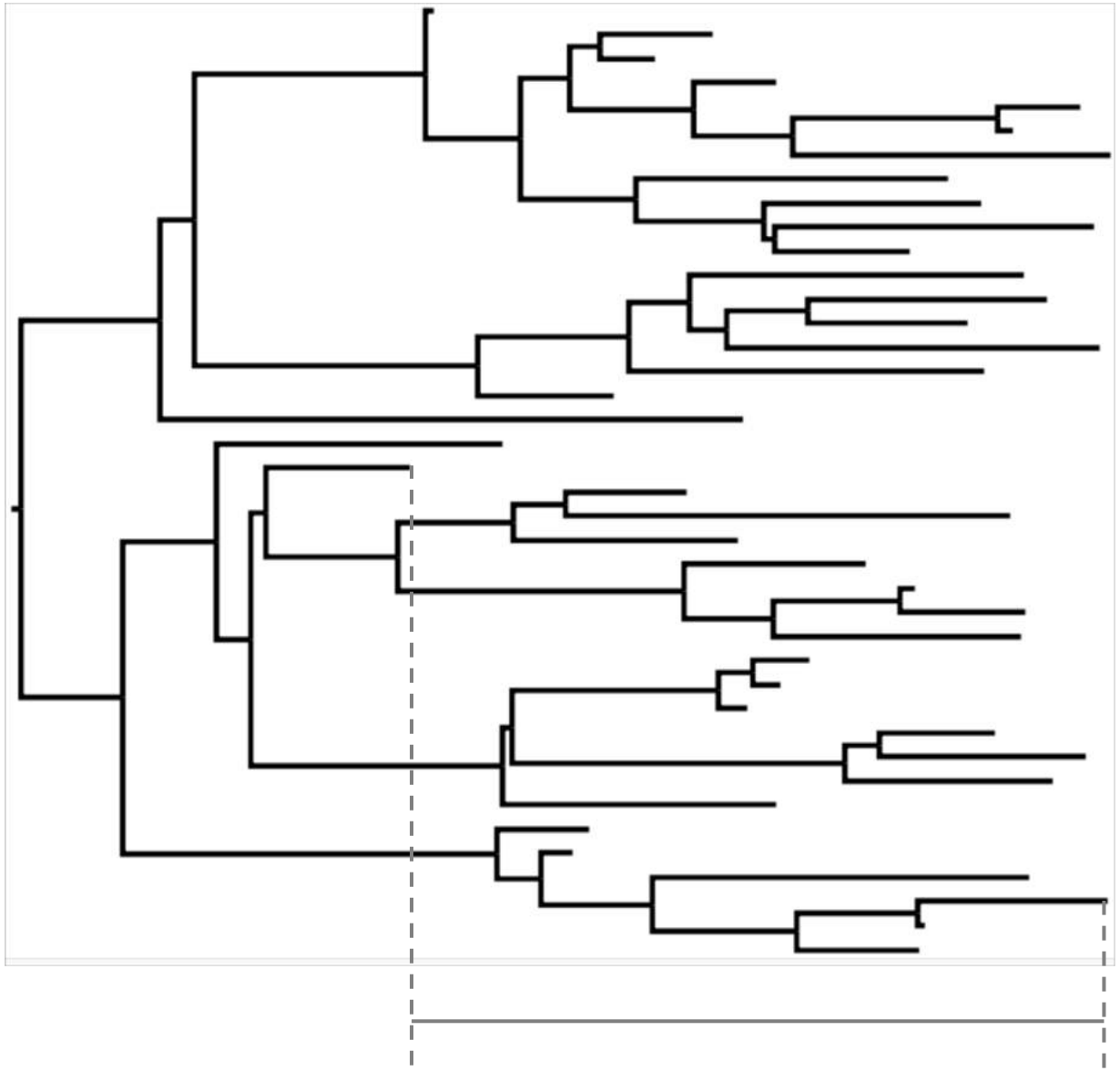
Number of tips

Birth-death constant sampling



10 Months

Birth-death skyline- variable sampling

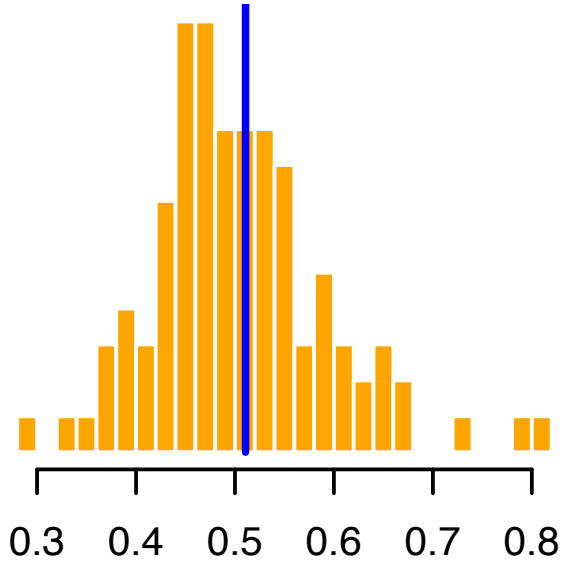


8 Months

Age range of tips

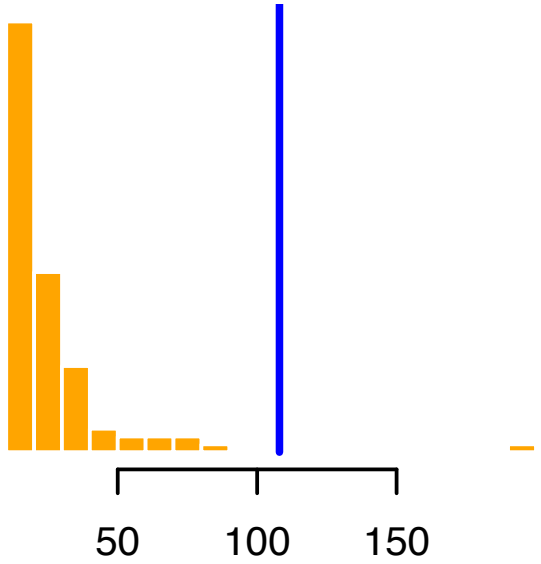
Birth-death constant sampling

Root height



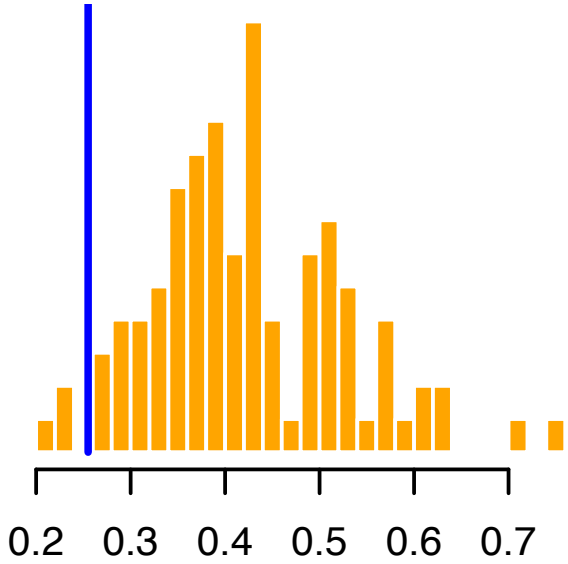
$P_B = 0.58$

Number of tips

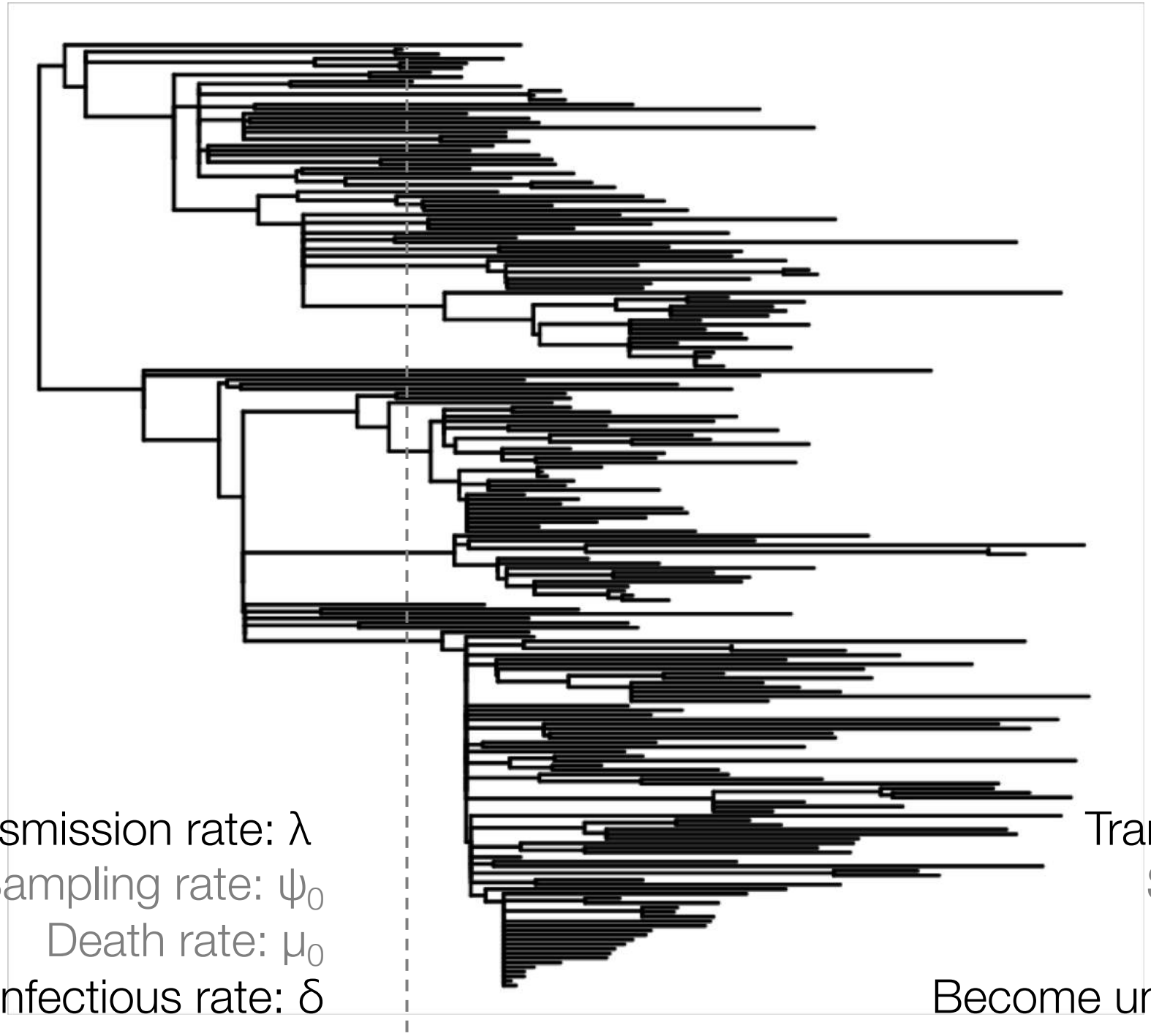


$P_B = 0.99$

Range of ages of tips



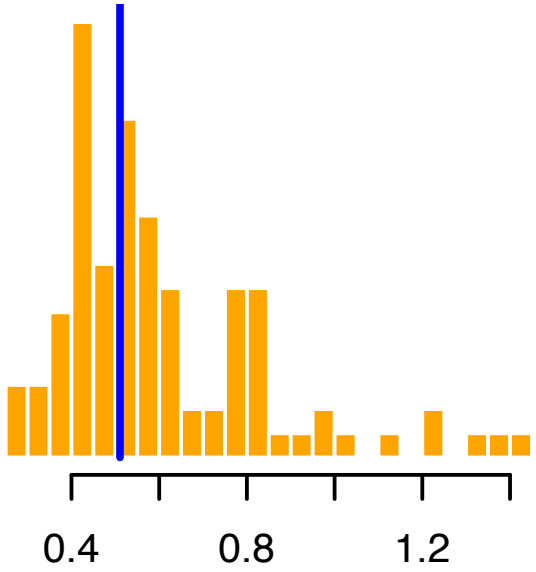
$P_B = 0.03$



Transmission rate: λ
Sampling rate: ψ_0
Death rate: μ_0
Become uninfectious rate: δ

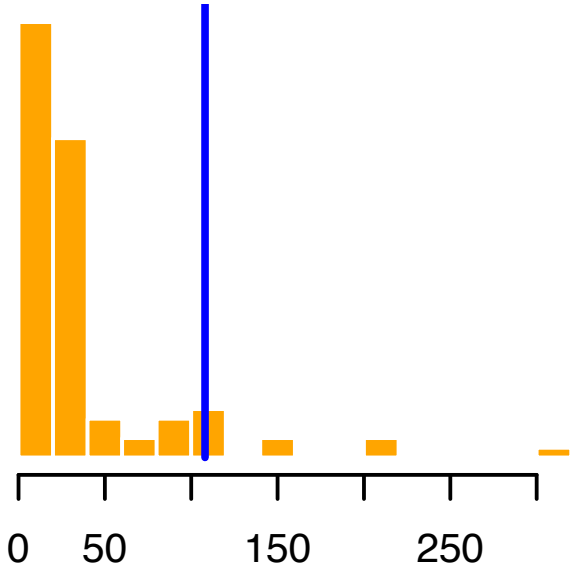
Transmission rate: λ
Sampling rate: ψ_1
Death rate: μ_1
Become uninfectious rate: δ

Root height



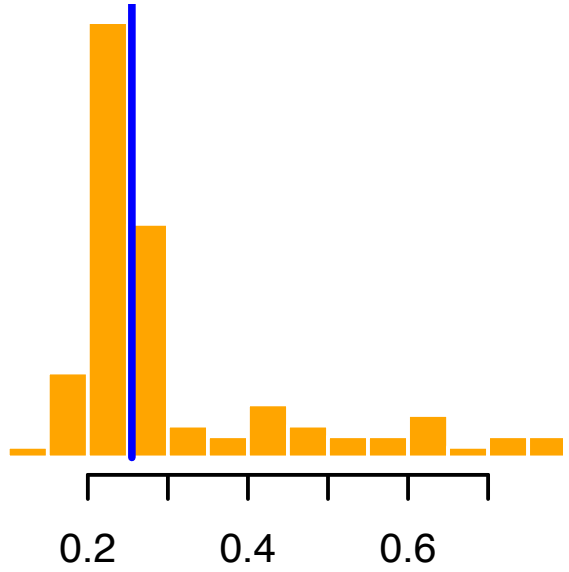
$P_B = 0.42$

Number of tips



$P_B = 0.91$

Range of ages of tips



$P_B = 0.61$

Final considerations

- Test statistics can be designed using simulations and should assess some expectation or assumption of the model.
- Ideally, test statistics should be a function of the data (certainly not sufficient statistics).
- Comparing the prior and posterior is key in Bayesian model adequacy.
- Model adequacy can be useful for model improvement and for understanding the reliability of the inferences.