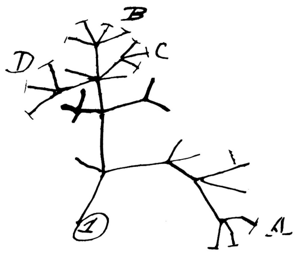
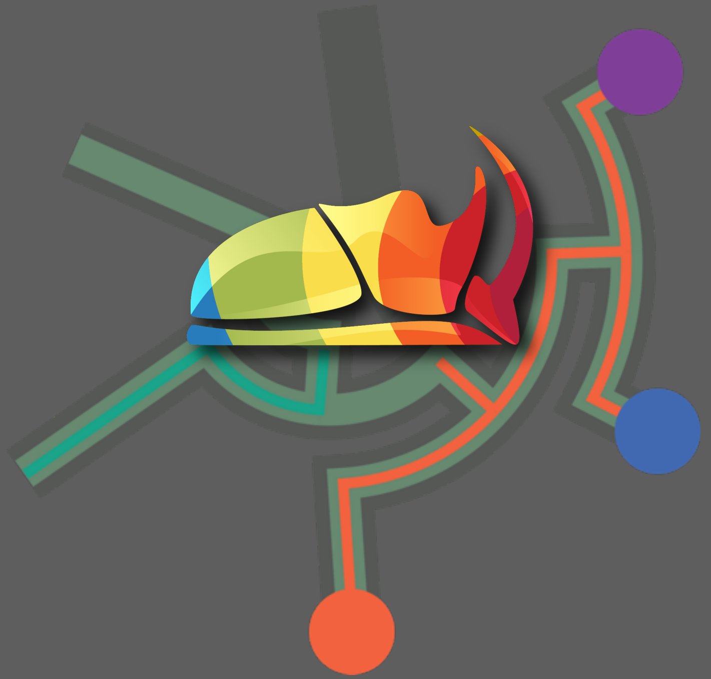


IPS-164 INTRODUCTION TO PHYLOGENETICS 2022

Lecture 11
Estimating Divergence Time



Sergei Tarasov

Beetle curator & Docent
Finnish Museum of Natural History, University of Helsinki

- @tarasov_sergio
- sergei.tarasov@helsinki.fi
- https://www.tarasovlab.com

PLAN OF THE TODAY’S LECTURE

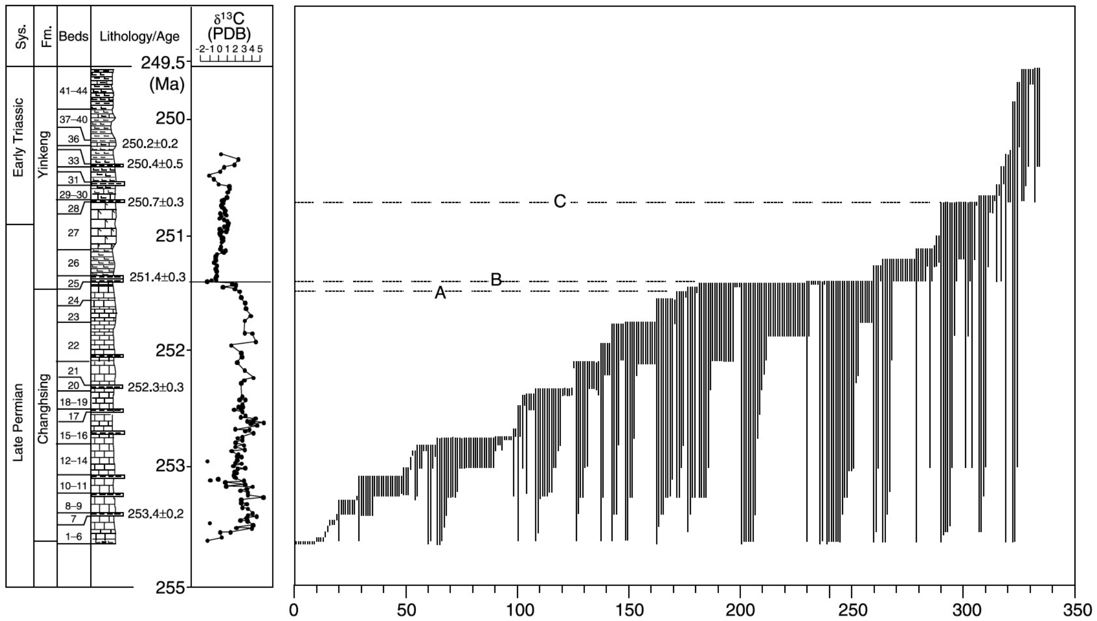
1. Main methods for divergence time estimation

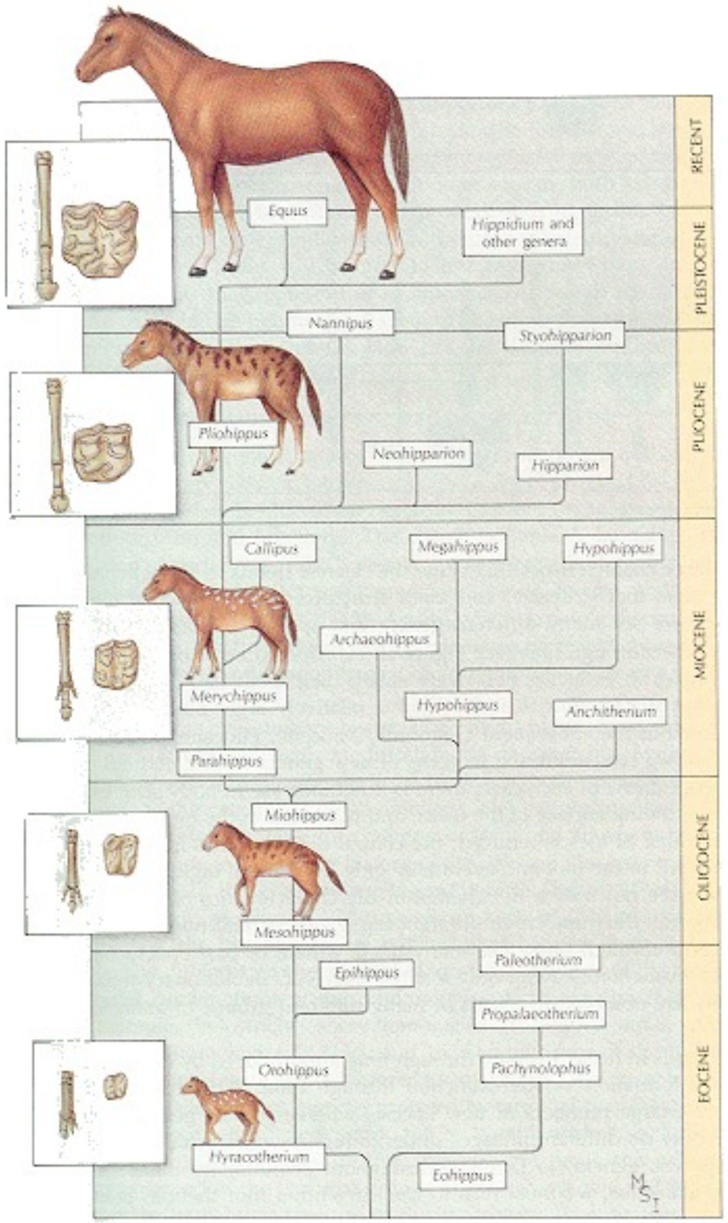
2. Comparative phylogenetics (to be continued on Monday)

Dating Phylogenetic Trees

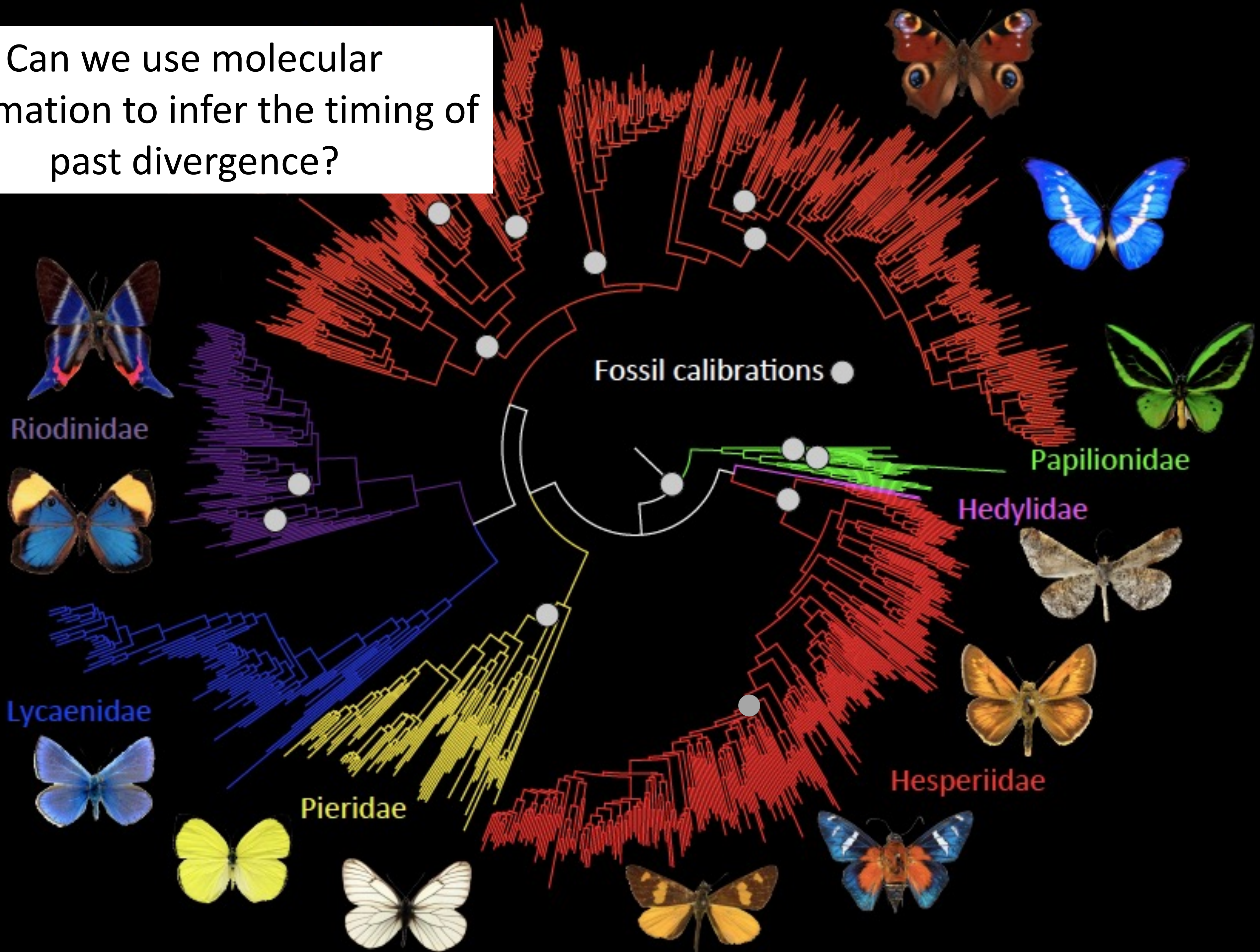
Many slides below are from Niklas Wahlberg

The fossil record is the direct evidence of past events and the time at which they occurred

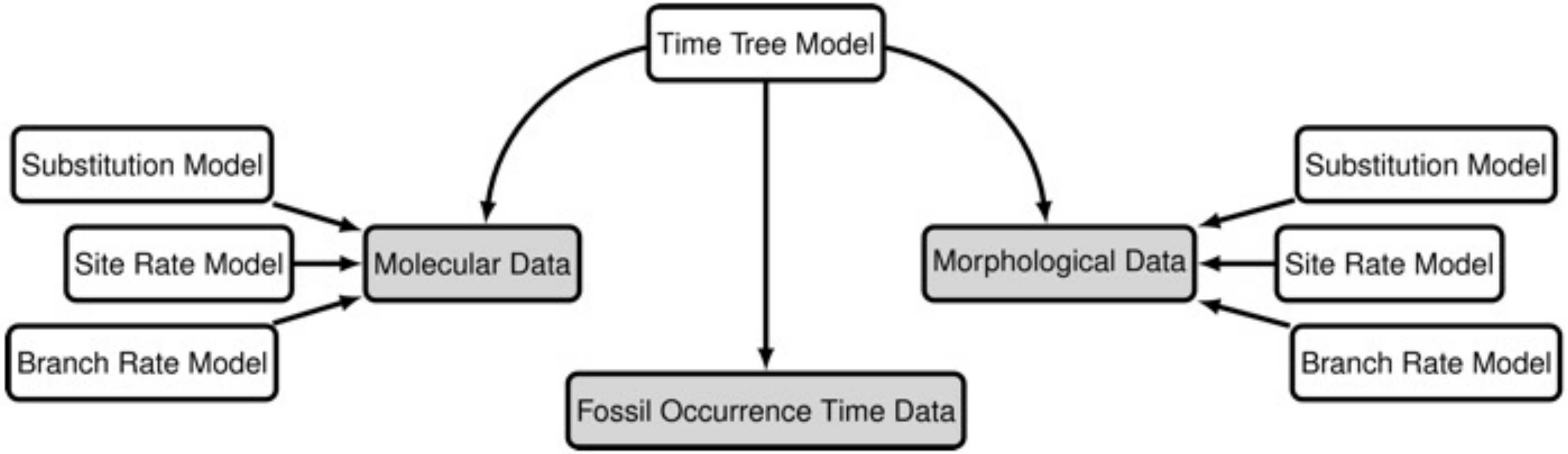




Can we use molecular information to infer the timing of past divergence?

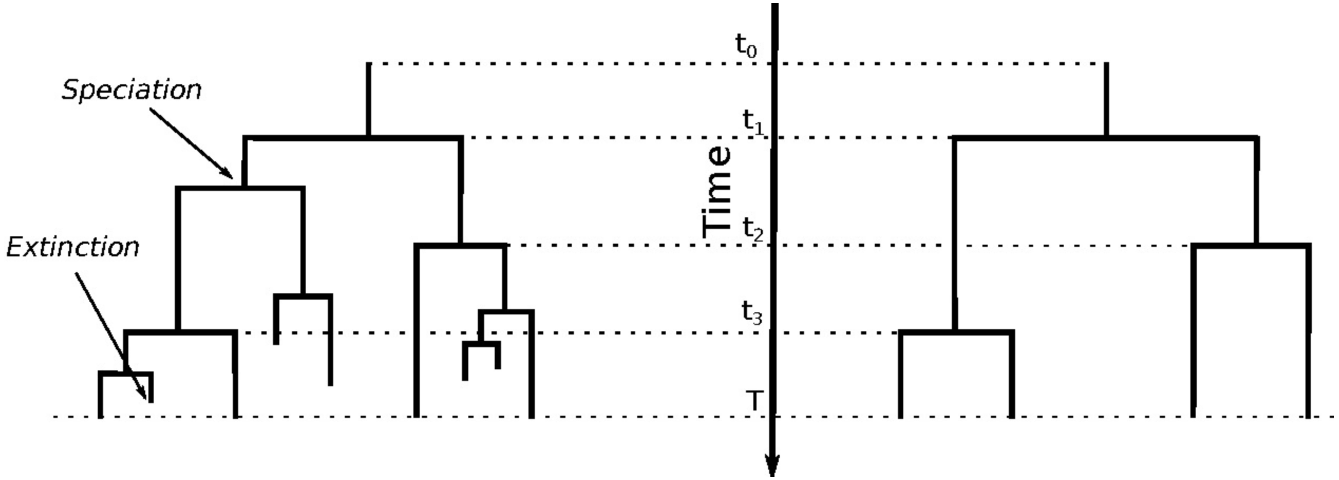


The ultimate aim is the "combined-evidence" analysis



https://revbayes.github.io/

Tree generating process



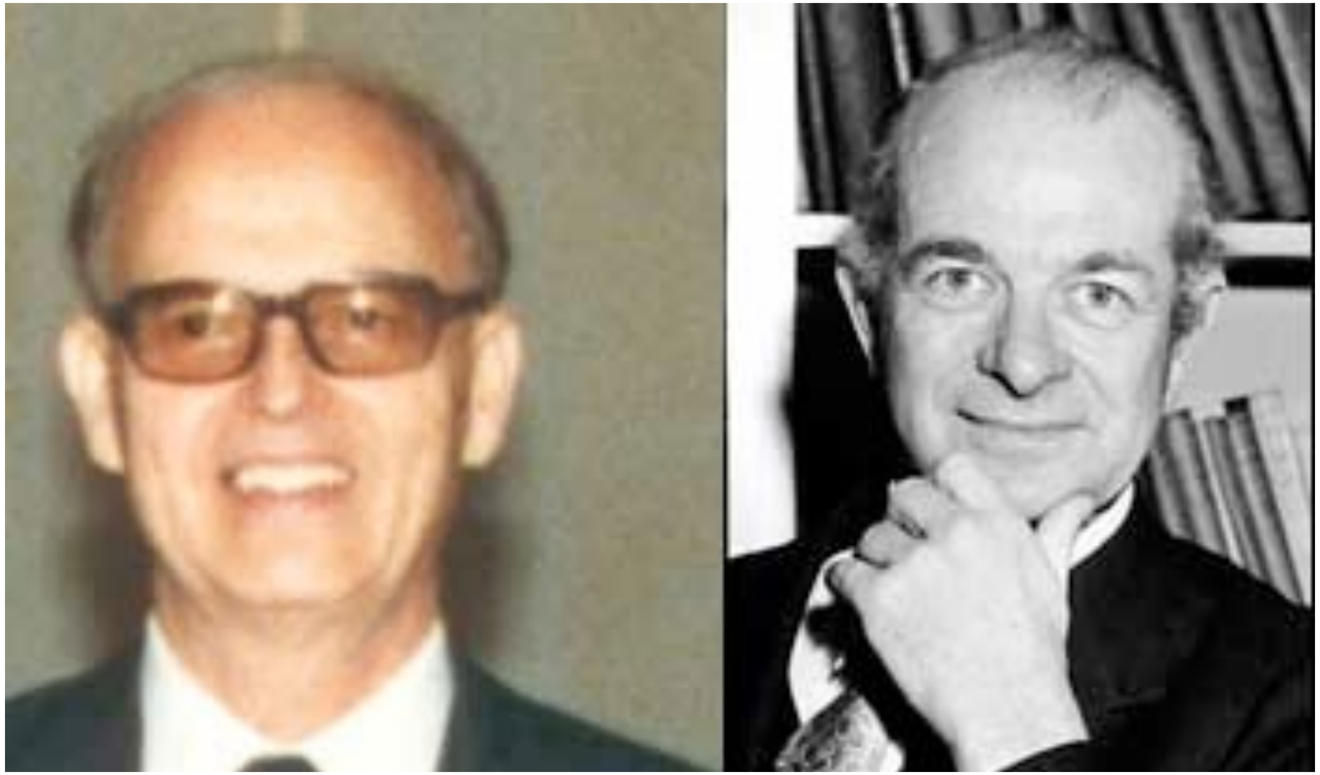
”Real” process

Observable process

The Molecular Clock

Going back to ancient times

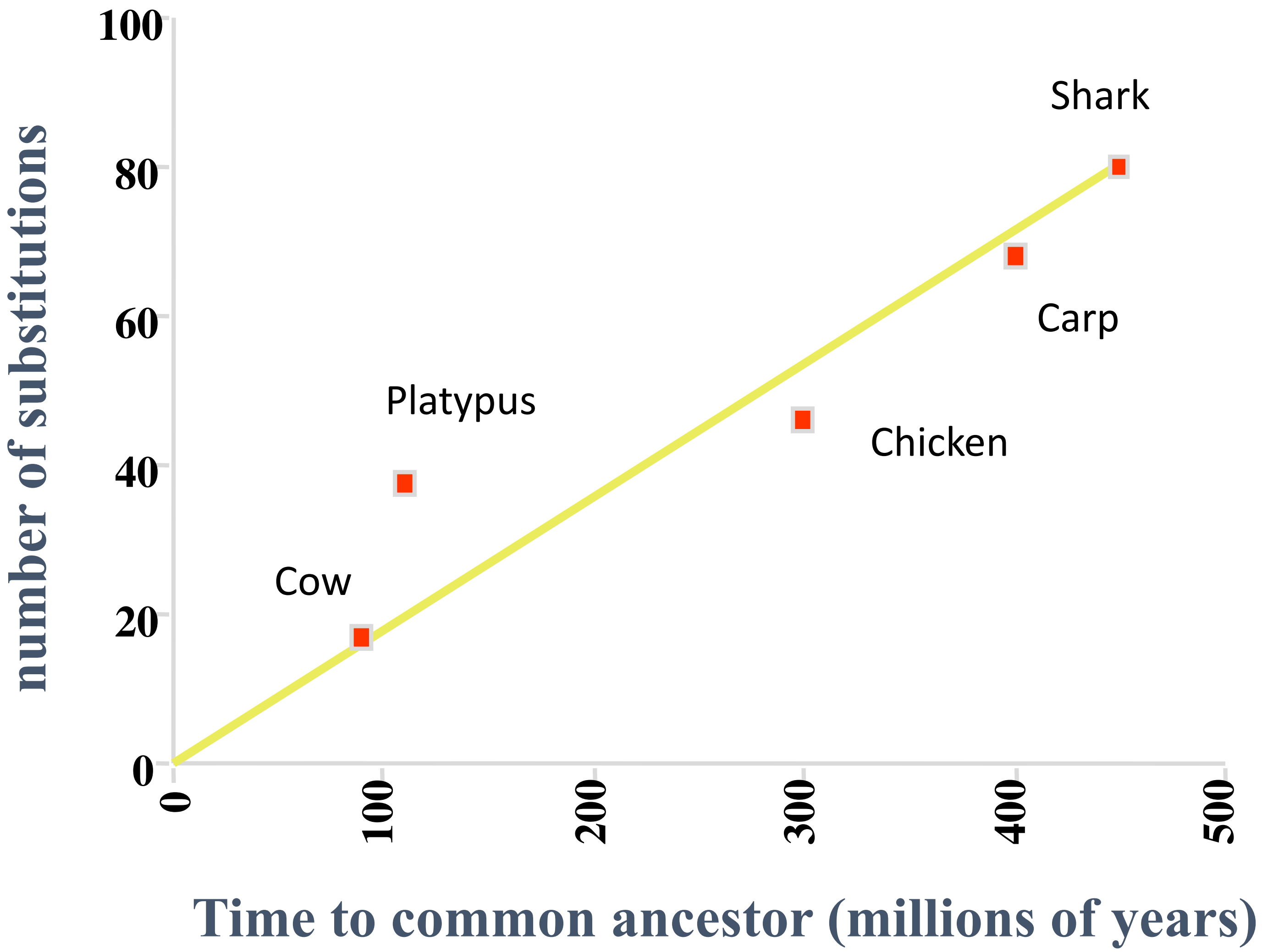
Is there a molecular clock?



- The idea of a molecular clock was initially suggested by Zuckerkandl and Pauling in 1962 and 1965

The molecular clock for alpha-globin:

Each point represents the number of substitutions separating each animal from humans

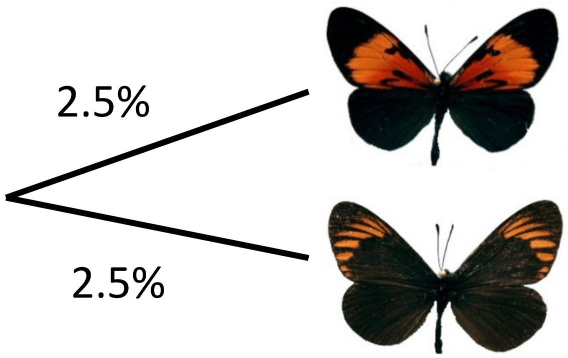


Is there a molecular clock?

- The idea of a molecular clock was initially suggested by Zuckerkandl and Pauling in 1962 and 1965
- They noted that rates of amino acid replacements in animal haemoglobins were roughly proportional to time - as judged against the fossil record
- This implies the existence of a sort of molecular clock ticking faster or slower for different genes but at a more or less constant rate for genes among different lineages

The molecular clock hypothesis

- Assumes an equal rate of molecular evolution over time



- A 5% difference between species means they have each diverged 2.5% since their common ancestor
- If a fossil or other evidence will let us calibrate this clock we can convert % difference to years

Assumptions of a perfect clock

- Molecular change is a linear function of time with substitutions accumulating following a Poisson distribution - any variation will be stochastic [imagine 1 substitution / million yrs]
- Rate of change is equal across all sites and lineages
- The phylogeny can be estimated without error

Assumptions of a perfect clock (cont.)

- The number of substitutions along each lineage can be estimated without error
- Calibration dates for all times of divergence used to calculate the rate of the molecular clock are known without error
- A regression of time on number of substitutions can be conducted without error

Dating with a molecular clock

- “Universal Molecular Clocks”
- Calibrations proposed for various taxa / genes
- eg. mtDNA molecular clock of animals
  - ~ 2% sequence divergence per million years for vertebrates
  - ~ 1% sequence divergence per million years for invertebrates

There is no universal molecular clock

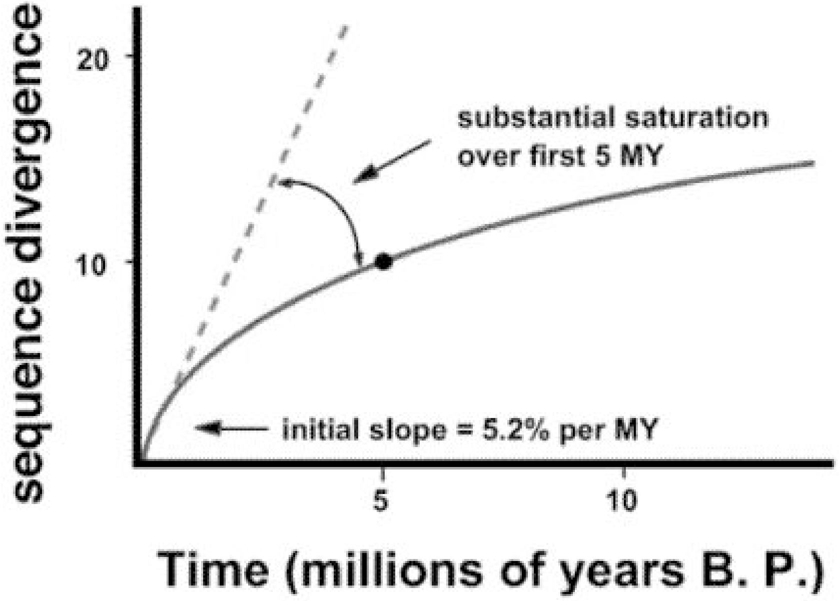
- The initial proposal saw the clock as a Poisson process with a constant rate
- Now known to be more complex - differences in rates occur for:
  - different sites in a molecule
  - different genes
  - different regions of genomes
  - different genomes in the same cell
  - different taxonomic groups for the same gene
- There is no universal molecular clock!!!

Problems

Saturation

Saturation problems

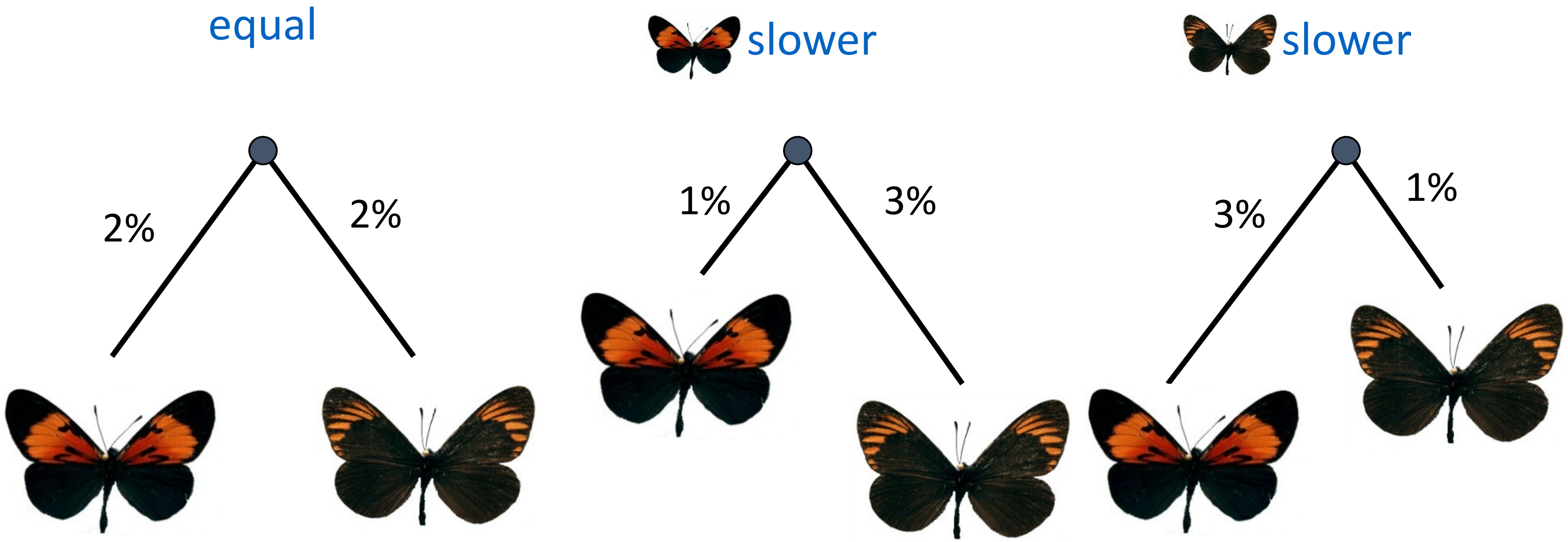
Saturation is the result of multiple substitutions in sequences. So, the apparent sequence divergence rate is lower than the actual divergence that has occurred.



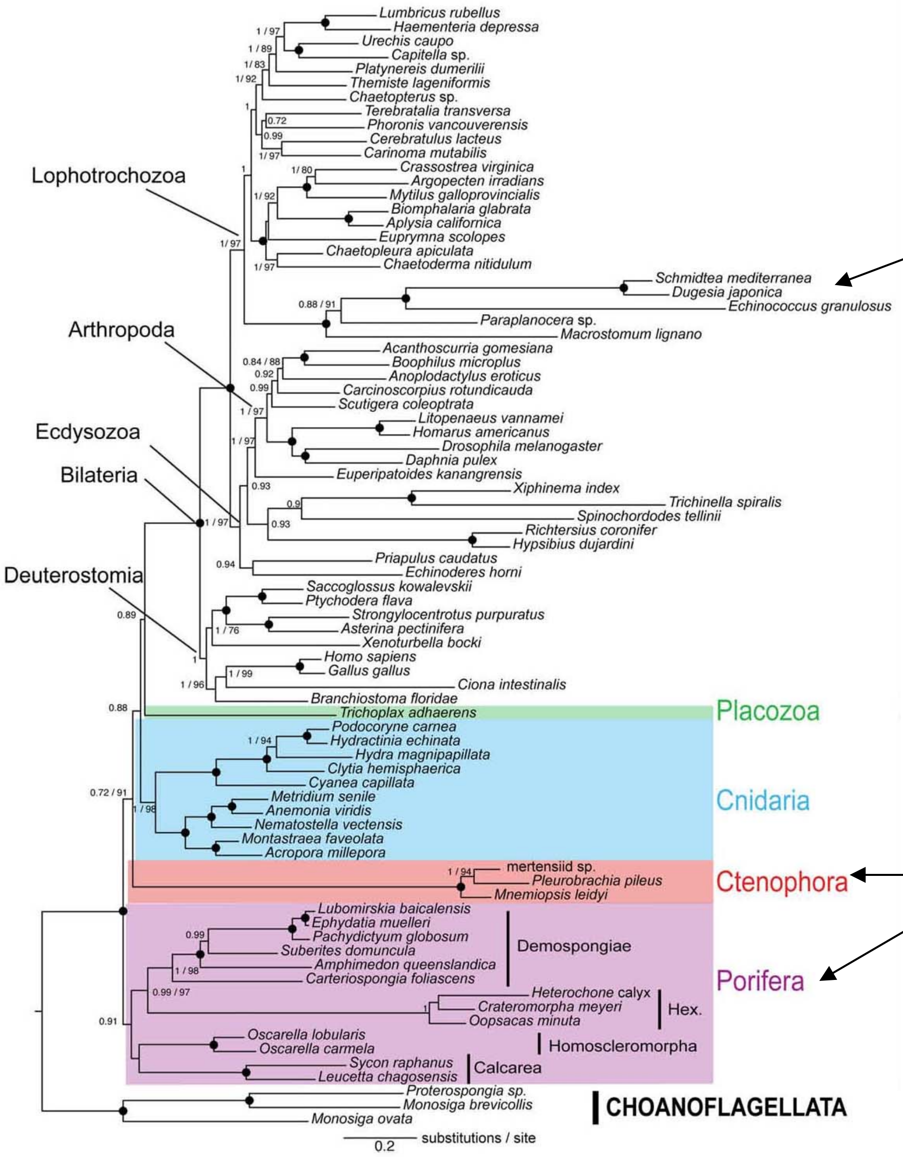
Problems

- Saturation
- Rate Heterogeneity - violation of homogeneity

No universal molecular clock



Molecular distance from to is the same in all cases



Long branches near the tip of a tree are probably “long” due to an increase in rate of change (not more time)

Long branches near the base of a tree are probably “long” due to time (not a faster rate of change)

Pick et al (2010) MBE 27:1983-1987

Teasing apart RATE and TIME

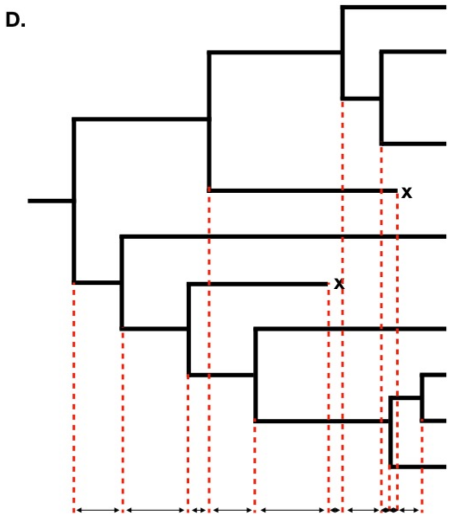
- Branch lengths are proportional to:

Branch Length = RATE * TIME

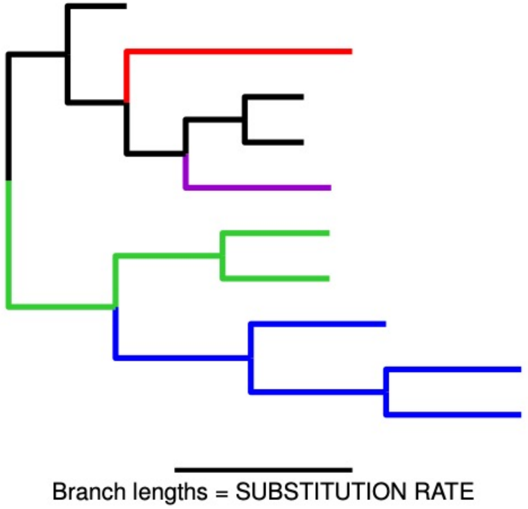
- If rates are constant then lengths are proportional to time
- If rates are not constant then *we have a hard time relating branch lengths to time*

Main components of divergence time models

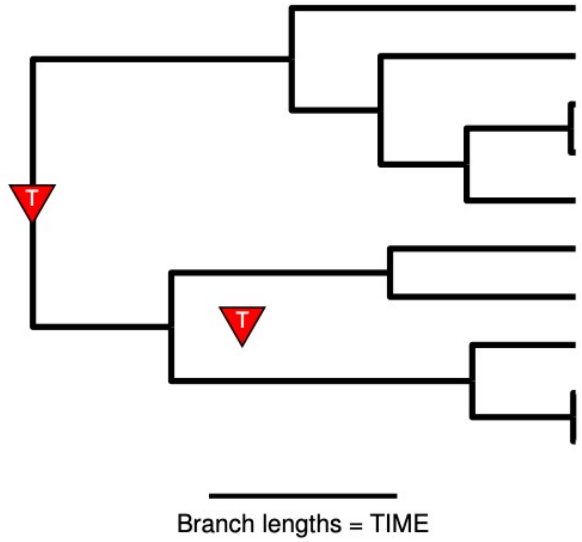
Tree birth-death model



Clock model
(rate variation across lineages)



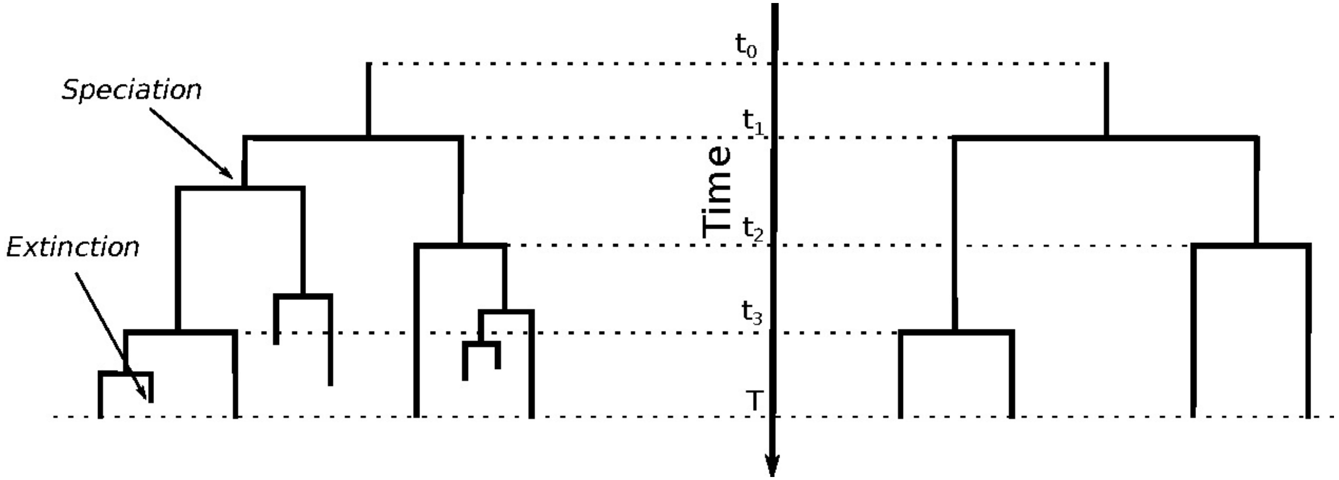
Calibration points



http://treethinkers.org/wp-content/uploads/2014/08/DivTime_BEAST2_tutorial_2014.pdf
https://lukejharmon.github.io/pcm/chapter10_birthdeath/

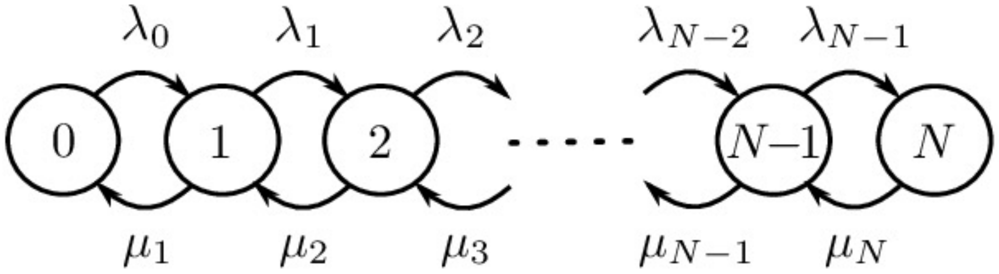
Tree birth-death model

Tree generating process



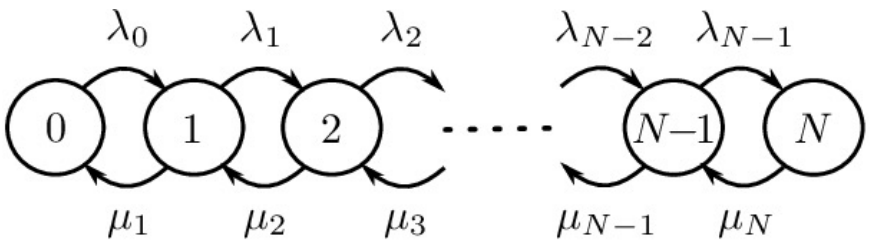
”Real” process

Observable process

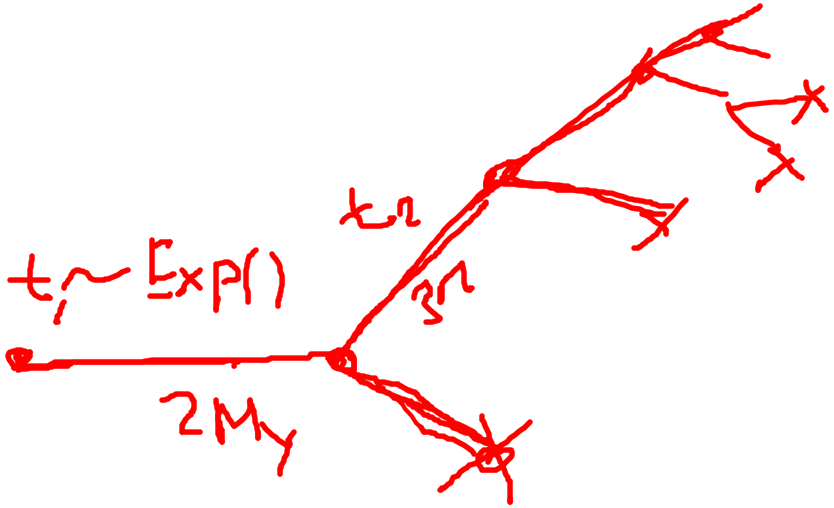


- Birth-death Markov model
- Two components: speciation and extinction

Modeling the tree generating process



$$P = \begin{array}{cccccc} & 1 & 2 & 3 & & \\ 1 & -\mu_0-\lambda_0 & \lambda_0 & 0 & 0 & \square \\ & \mu_1 & -\mu_1-\lambda_1 & \lambda_1 & 0 & \square \\ & 0 & \mu_2 & -\mu_2-\lambda_2 & \lambda_2 & \square \\ & 0 & 0 & \mu_3 & -\mu_3-\lambda_3 & \square \\ & \square & \square & \square & \square & \square \end{array}$$



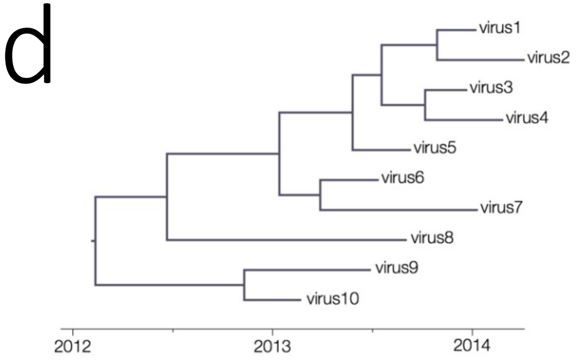
- Speciation rate: $\boldsymbol{\lambda}$
- Extinction rate: $\boldsymbol{\mu}$
- Waiting time:
  - nothing happens: $-\boldsymbol{\mu}-\boldsymbol{\lambda}$
  - so, the waiting time is *Exp($\boldsymbol{\mu}+\boldsymbol{\lambda}$)*

$$\underline{SP} = \frac{\lambda}{\lambda+\mu} \qquad ex = \frac{\mu}{\lambda+\mu}$$

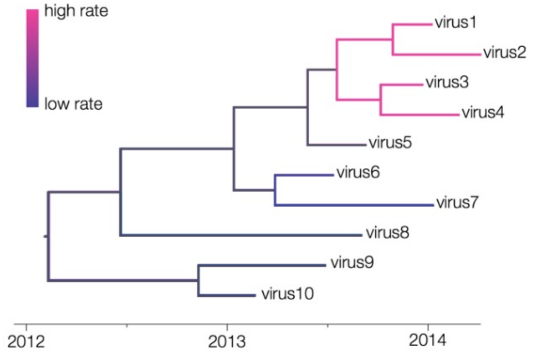
Molecular Clock models

Molecular clocks can be relaxed

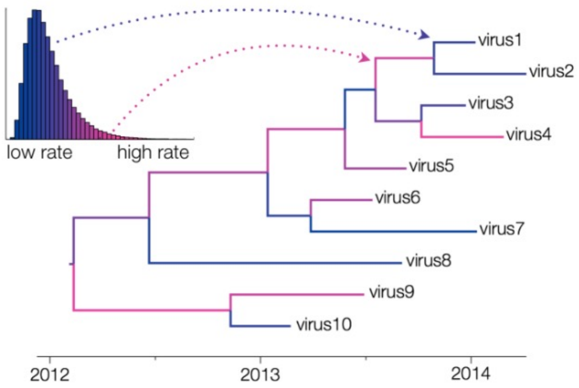
- Strict or ”global” clock
  - Many programs/methods/algorithms



- Local clocks
  - Maximum Likelihood (PAML, QDate)
  - Mean path length (Pathd8)

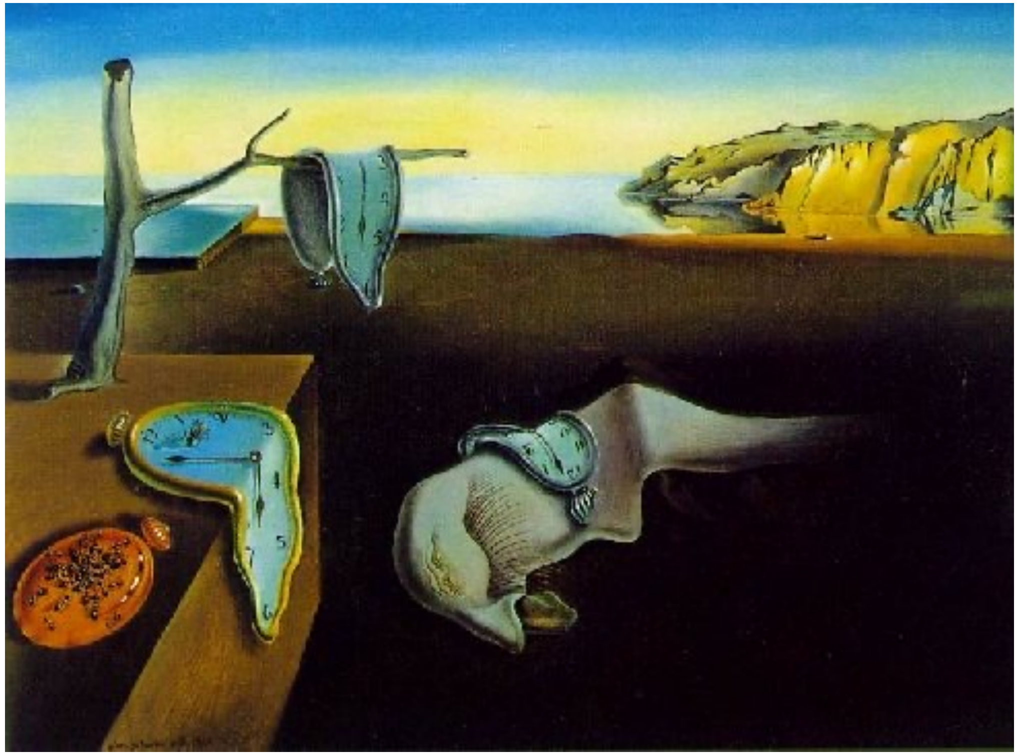


- Relaxed clocks
  - Non-parametric rate smoothing (r8s)
  - Penalized likelihood (r8s)
  - Bayesian, fixed tree (multidivtime, PhyBayes)
  - Bayesian, tree co-estimated (BEAST, MrBayes, RevBayes)



What is a relaxed clock?

- Strict clock: rate identical in all branches
- Relaxed clock: rate allowed to vary among branches
  - Autocorrelated relaxed clock: rates in adjacent branches are related
  - Uncorrelated relaxed clock: rates identically and independently distributed among branches



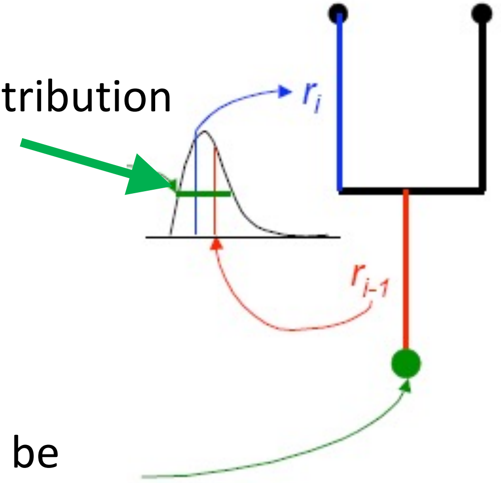
Autocorrelated relaxed clocks

- Fixed topologies are input!
- Treat substitution rate as a heritable trait, so that it can ‘evolve’ through the tree
- Rate is assumed to be tied to:
  - Life history traits (e.g., generation time, population size, body size)
  - Cellular/biochemical environment
- Available in *r8s, multidivtime, PhyBayes, BEAST, PAML, RevBayes*

Modeling autocorrelation

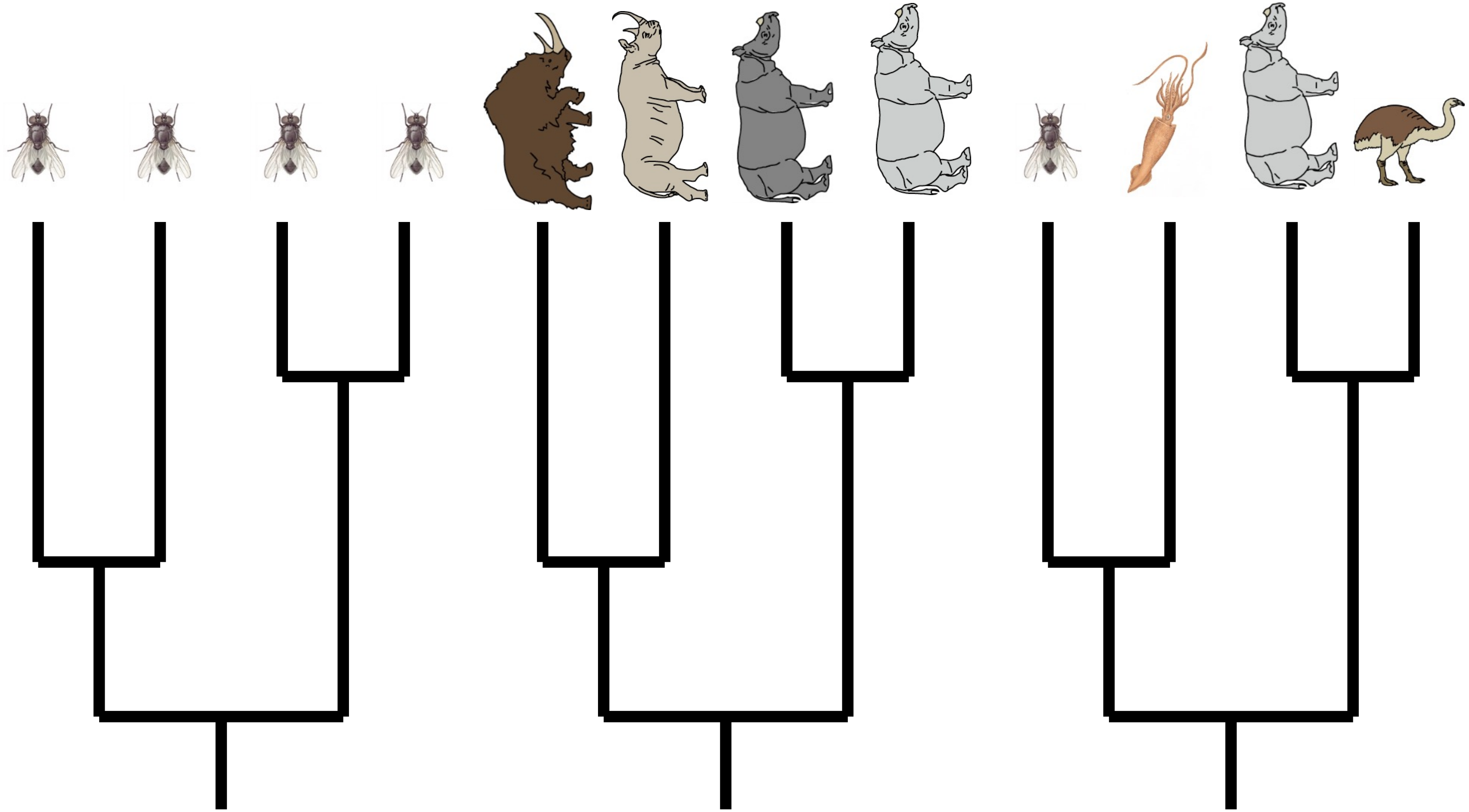
- Model of autocorrelated rate change used to describe prior distribution of rates
- Available in *BEAST*
- Lognormal
  - $log(\mathbf{r}_i) \sim N(log(\mathbf{r}_i\mathbf{-1}), \mathbf{vt})$

v controls the s.d. of the distribution



Further assumption needs to be made about rate at the root

Autocorrelated relaxed clocks



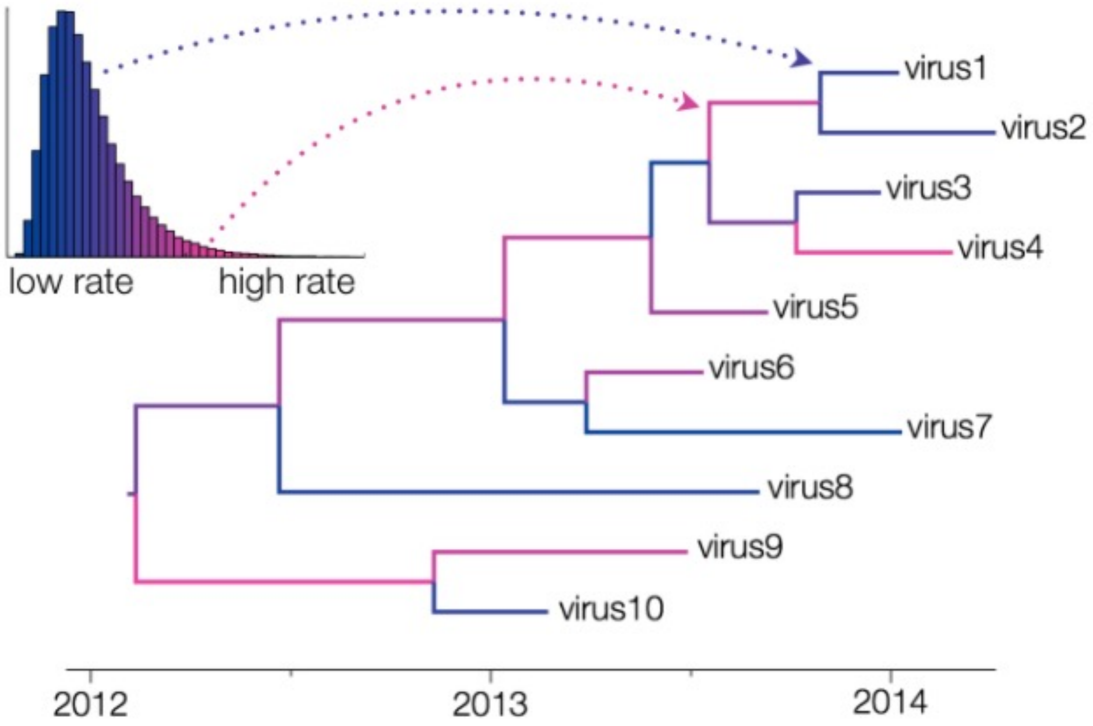
Very high autocorrelation

Moderate autocorrelation

Very low autocorrelation

Uncorrelated relaxed clocks

- Models available in *BEAST, RevBayes*
  - **Lognormal distribution**
    Most rates cluster around the mean
  - **Exponential distribution**
    Most rates are quite low



Lognormal uncorrelated relaxed clock

- In the uncorrelated lognormal relaxed clock, two statistics can be obtained:
  - **Coefficient of variation of rates**
    Measures the rate variation among branches
    A value of 0 indicates clocklike evolution
  - **Covariance of rates**
    Measures autocorrelation of rates between adjacent branches

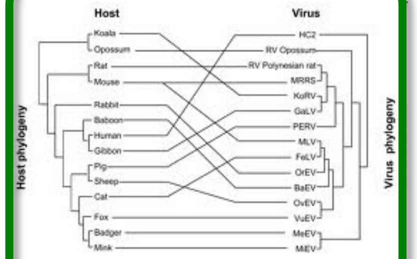
Calibration points

Separating rate and time

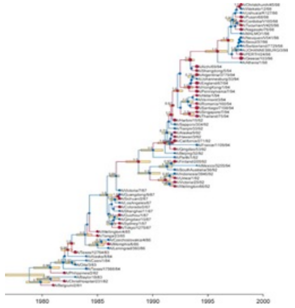
- Information about rate
  - Substitution rate obtained from an independent study
- Information about time – *prior information*:

Fossil record

Biogeography



Ecology

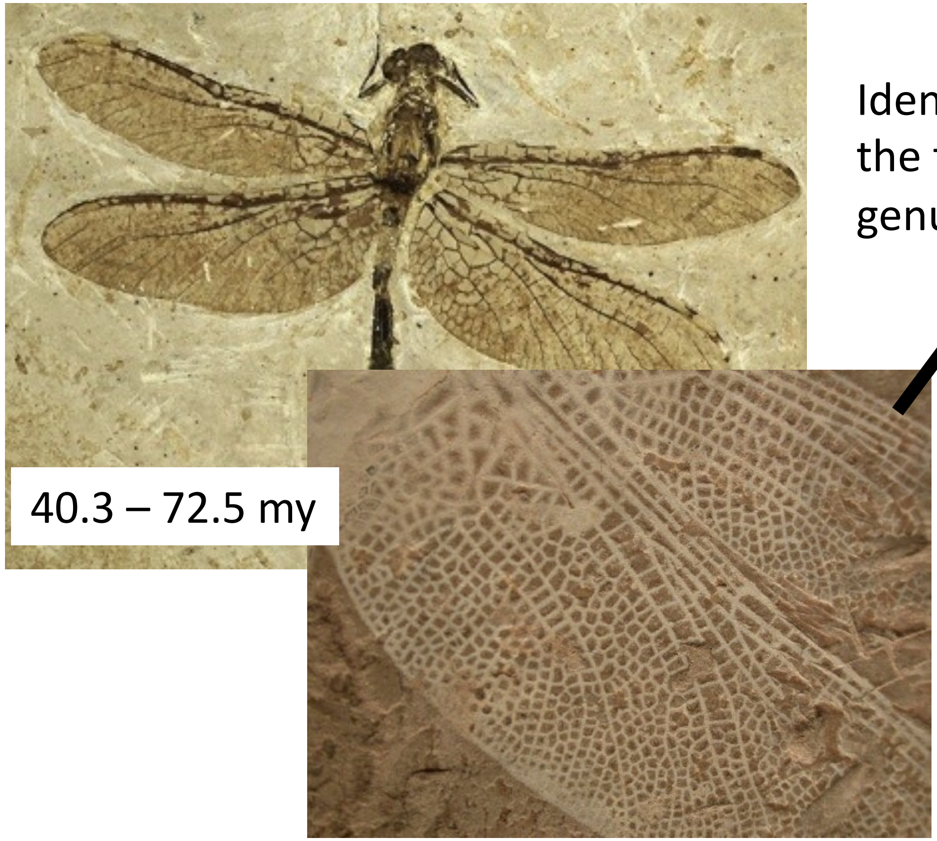


Sampling times

Calibration: Fossil record

- Fossil record provides minimum estimates of divergence times

Fossil record



40.3 – 72.5 my

Identified as belonging to the family Aeshnidae and genus *Aeshna*

informative

Calibration: Fossil record

- Fossil record provides minimum estimates of divergence times

Fossil record

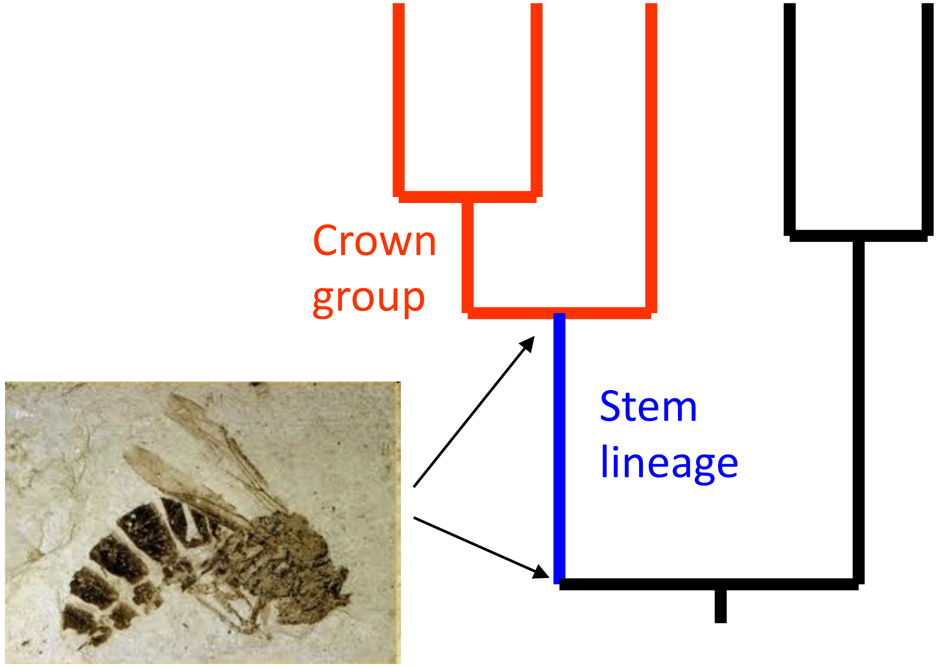
0 my

Identified as belonging to the family Phasianidae, to genus *Gallus*, to the species *Gallus gallus domesticus*

Minimum age for the birds…
BUT not informative

Problems with fossils

- Incompleteness of fossil record
- Identification
  - Species / Genus / Family?
- Position
  - Stem or crown?
- Which date?
  - Min / Mid / Max of Epoch?



Fossil from Eocene (55-33 Mya)

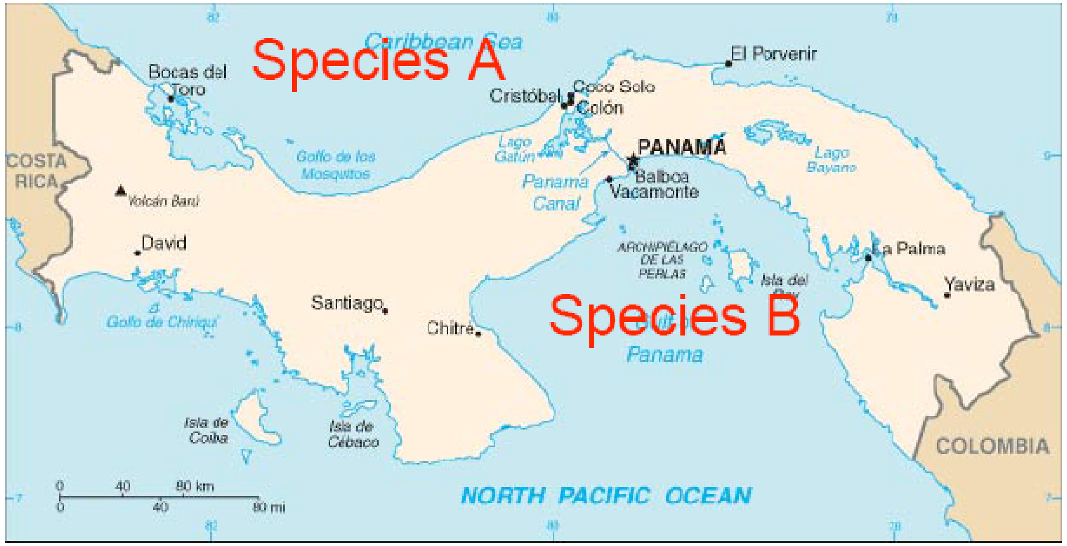
Fossil age errors

- Preservational bias
  - Hard parts
  - Environment, proximity to water bodies
  - Age
  - Sampling effort
- Taxonomic affinity
  - Fragmentary fossils
  - Extinct, stem lineages
- Stratigraphic and isotopic dating errors

Calibration: Biogeography

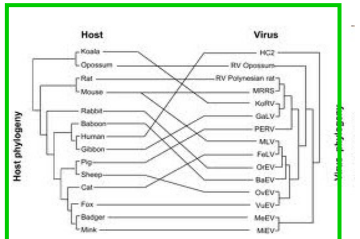
- Biogeographic events can provide maximum estimates of divergence times (NOT ALWAYS)

Biogeography

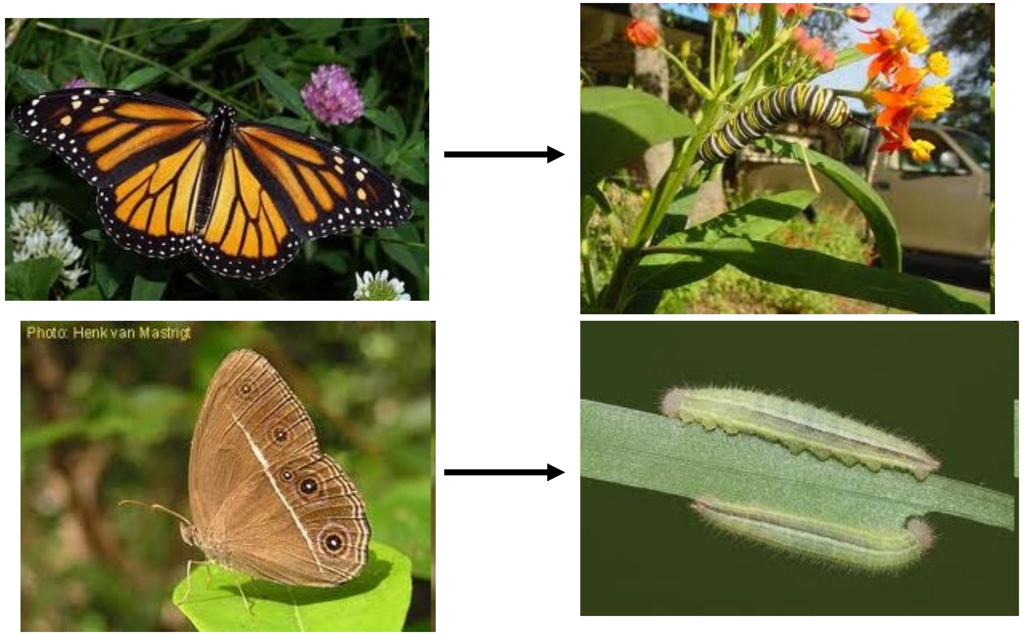


Calibration: Ecology

- Knowledge of tight ecological associations can be used to provide maximum estimates of divergence times (NOT ALWAYS)

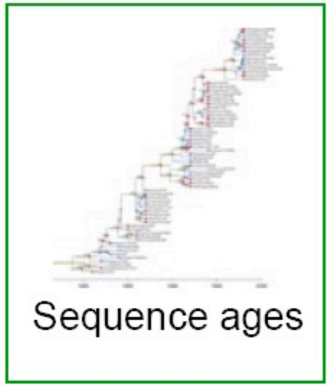


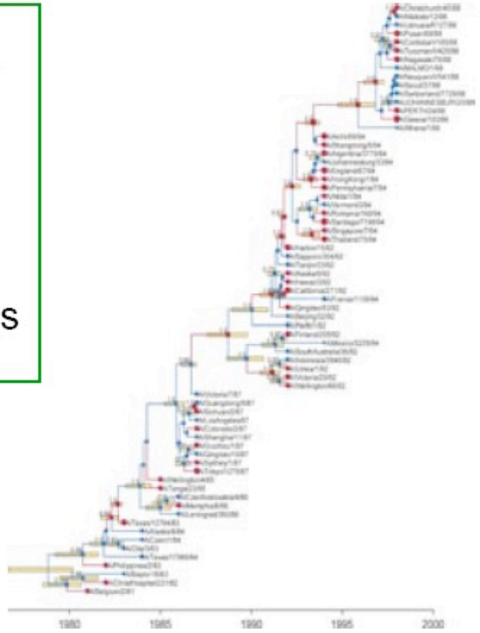
Ecology

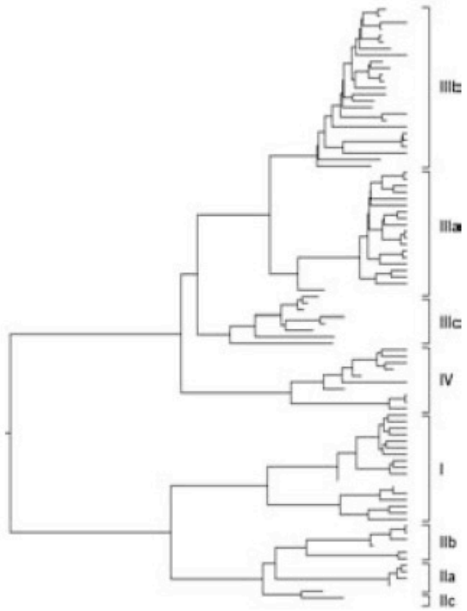


Calibration: Sequence ages

- Sequence ages provide sufficient age information for e.g. viruses







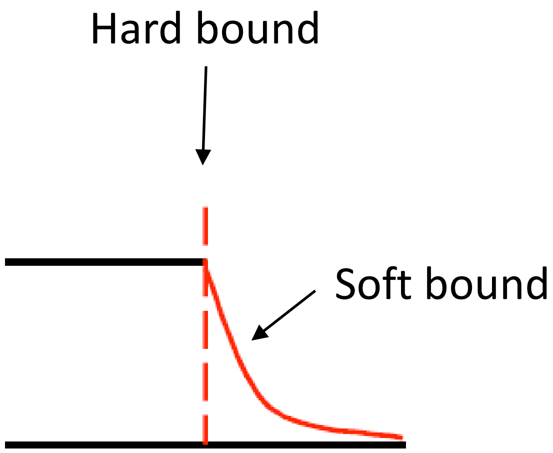
Modeling Calibration points

Calibration types

- Point calibrations
- Hard minimum/maximum bounds
- Parametric prior distributions
  - Normal distribution
  - Lognormal distribution
  - Exponential distribution

Hard/Soft Bounds

- Extension of hard bounds
- Soft:
  - Assign non-zero probability to values outside bound
  - Able to forgive calibration errors



Calibration in Bayesian framework

$$\overset{\text{posterior}}{f(\theta|D)} = \frac{\overset{\text{data}}{f(D|\theta)}\,\overset{\text{prior}}{f(\theta)}}{\int f(D|\theta)f(\theta)d\theta}$$

θ : model (substitution model(s), tree, etc)

prior: prior expectation we have for parameters of the model

For example: age of the nodes based on fossil information

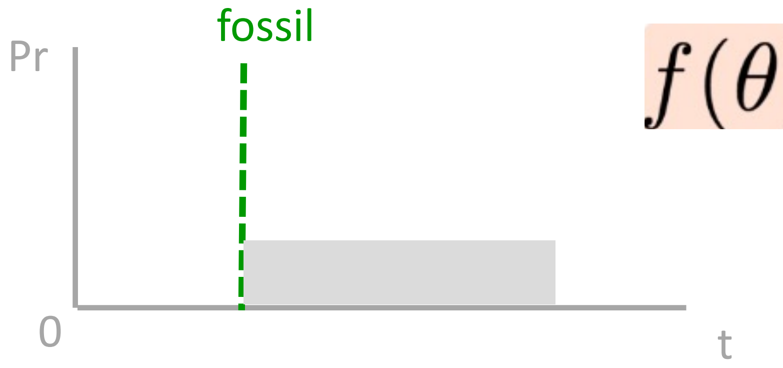
Calibration in Bayesian framework

posterior data prior

$$f(\theta|D) = \frac{f(D|\theta)f(\theta)}{\int f(D|\theta)f(\theta)d\theta}$$

For example: age of nodes based on fossil information

Prior probability distribution:



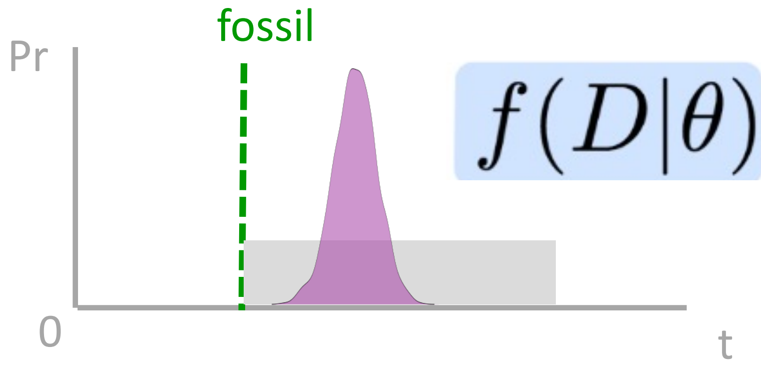
$f(\theta)$

Calibration in Bayesian framework

$$\overset{\text{posterior}}{f(\theta|D)} = \frac{\overset{\text{data}}{f(D|\theta)}\,\overset{\text{prior}}{f(\theta)}}{\int f(D|\theta)f(\theta)d\theta}$$

For example: age of nodes based on fossil information

Prior probability distribution:

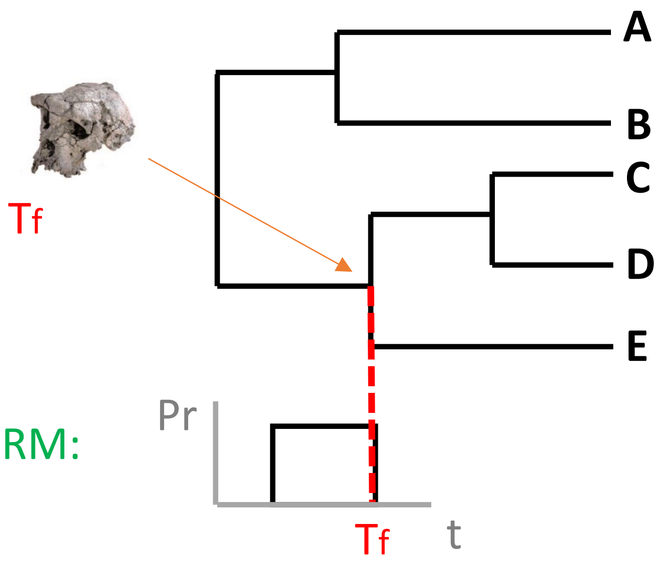


$$f(D|\theta)f(\theta)$$

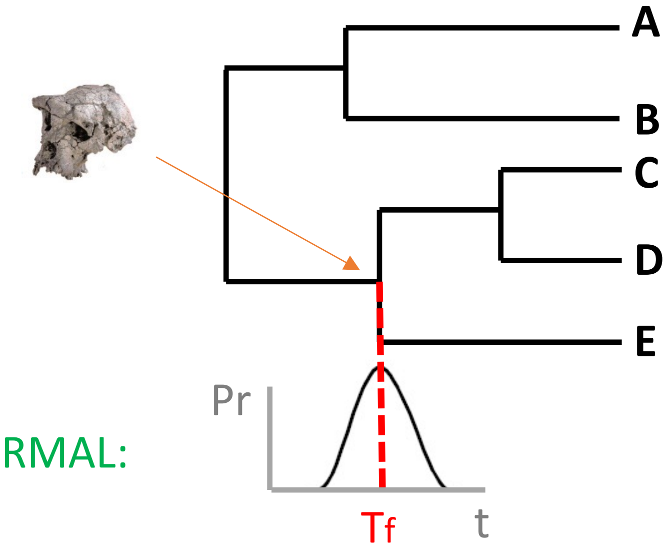
Prior distributions

fossil: Tf

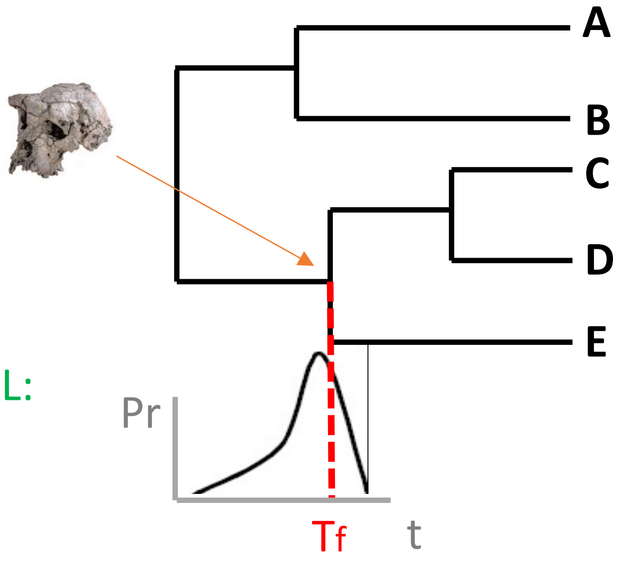
UNIFORM:



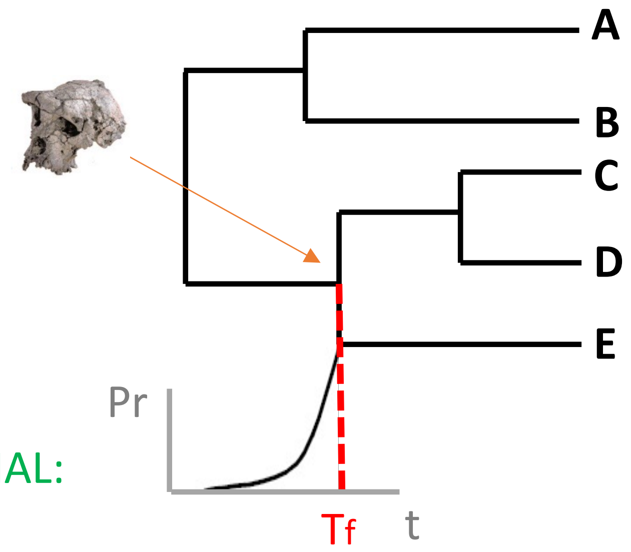
NORMAL:

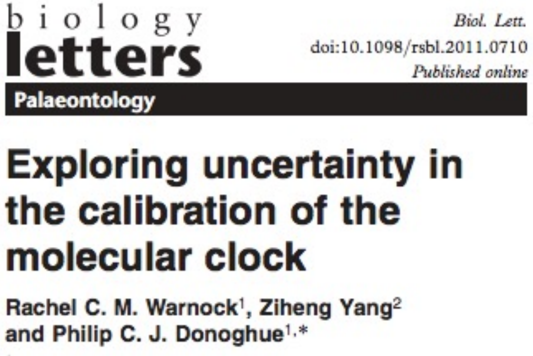


LOGNORMAL:



EXPONENTIAL:





biology letters

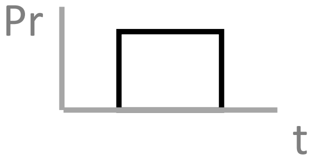
Biol. Lett.
doi:10.1098/rsbl.2011.0710
Published online

Palaeontology

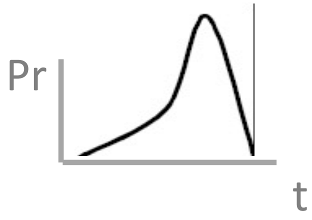
Exploring uncertainty in the calibration of the molecular clock

Rachel C. M. Warnock[1], Ziheng Yang[2] and Philip C. J. Donoghue[1,*]

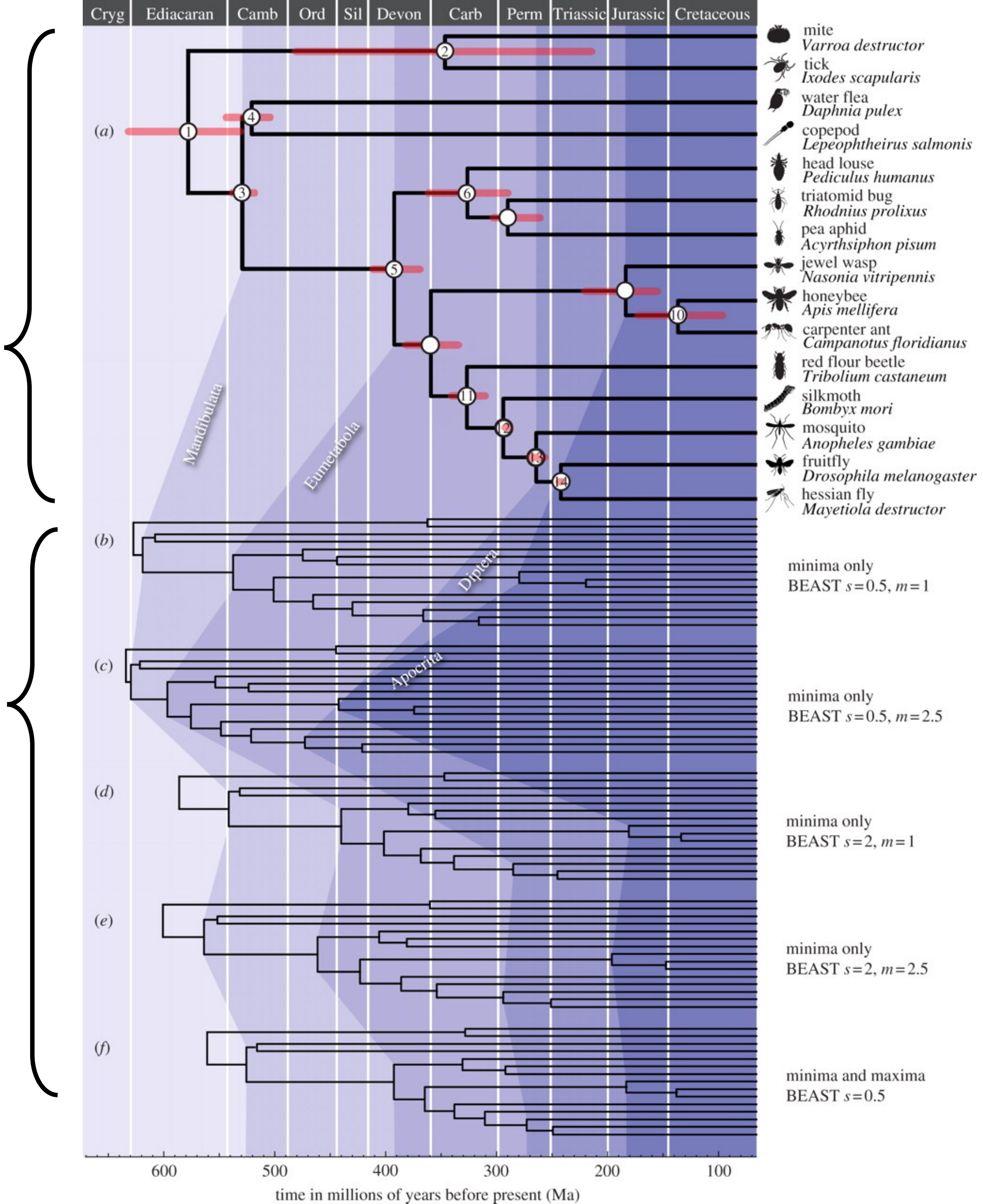
UNIFORM



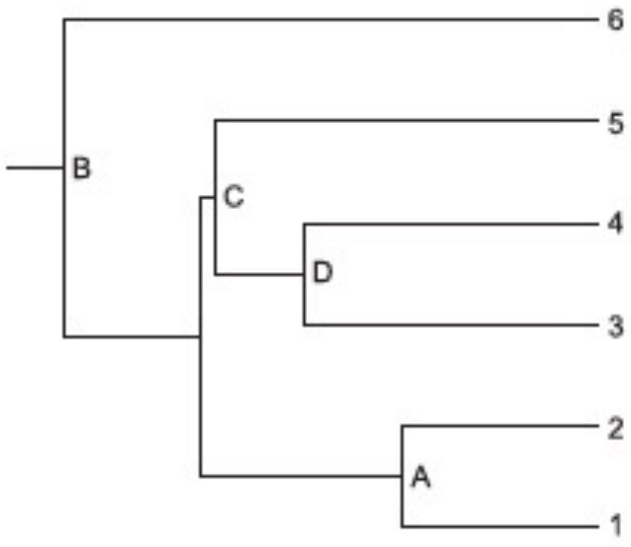
LOGNORMAL



m=mean
s=standard deviation

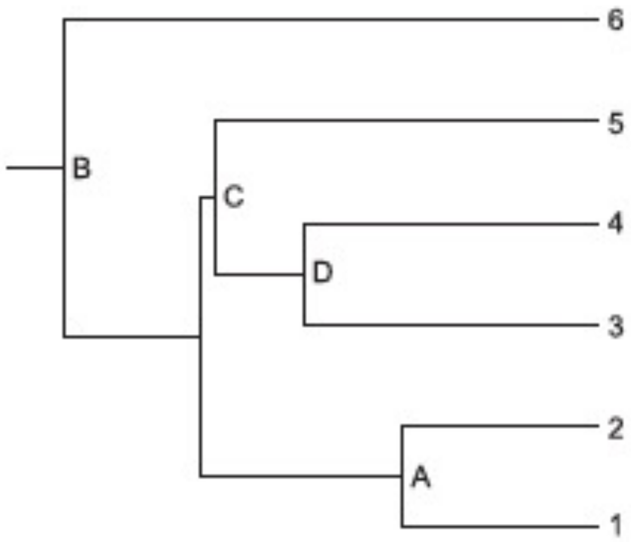


Multiple calibrations



- Molecular-clock estimates can be sensitive to the positions of the calibrations in the phylogenetic tree, especially when only a single or very few calibrations are available
- a small number of calibrations can lead to a biased estimate of the substitution rate if there is substantial among-lineage rate variation

Multiple calibrations



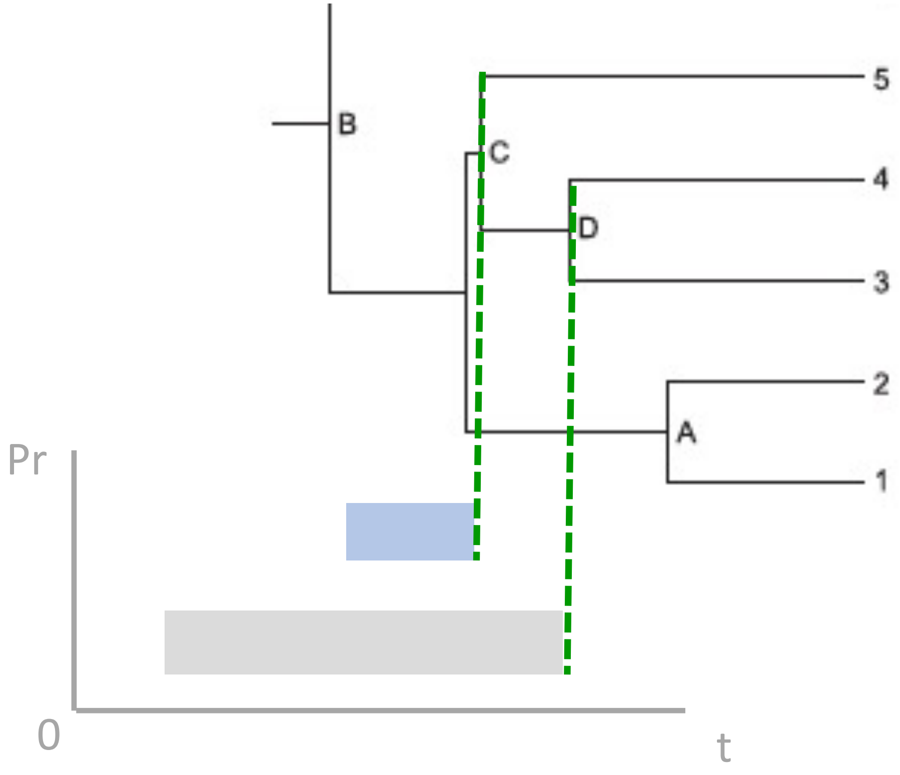
- substitution rate is primarily estimated from the branches between the calibrating nodes and the tips
- deeper calibrations capture a larger proportion of the overall genetic variation.

Multiple calibrations

- Be careful: priors interact with each others

- For example, node orders

D cannot be older than C



Multiple calibrations

- Be careful: priors interact with each others

- For example, node orders

- Marginal priors resulting from prior interactions can differ from the initial user prior
  - This can be visualized by removing the data and running the same analysis

$$f(\theta|D) = \frac{f(D|\theta)f(\theta)}{\int f(D|\theta)f(\theta)d\theta}$$

Multiple calibrations

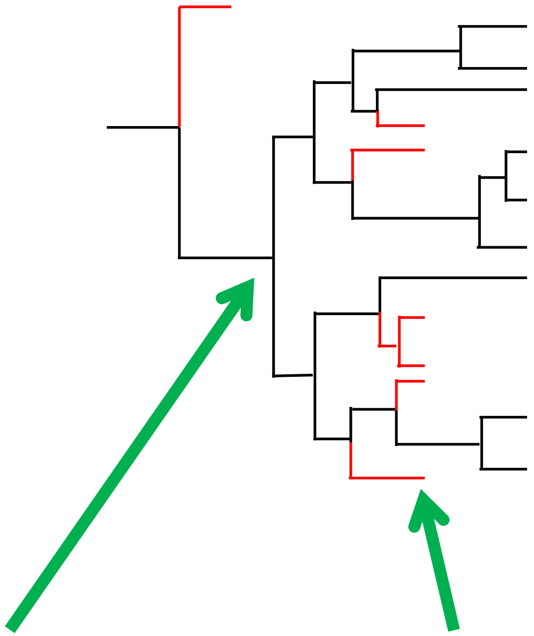
- Be careful: priors interact with each others

- For example, node orders

- Marginal priors resulting from prior interactions can differ from the initial user prior
  - This can be visualized by removing the data and running the same analysis

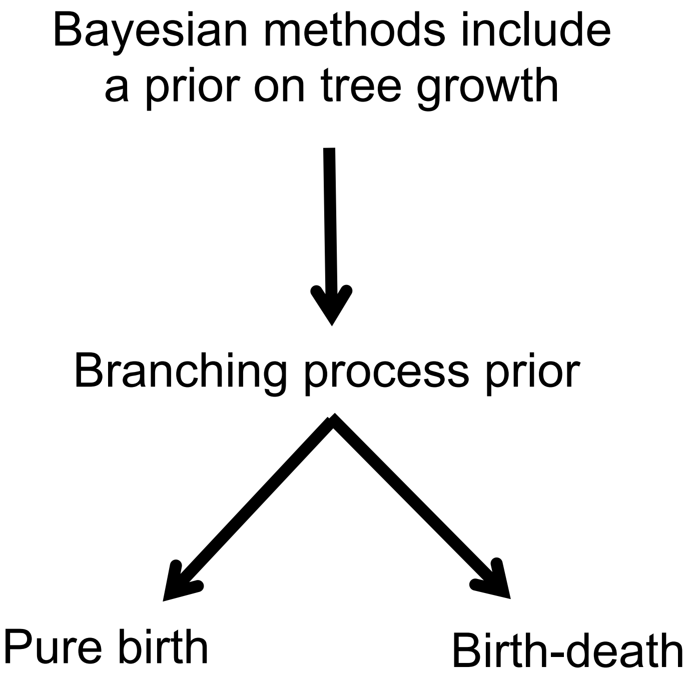
$$f(\theta|D) = \frac{f(\theta)}{\int f(D|\theta)f(\theta)d\theta}$$

Prior sensitivity



speciation process

extinction process



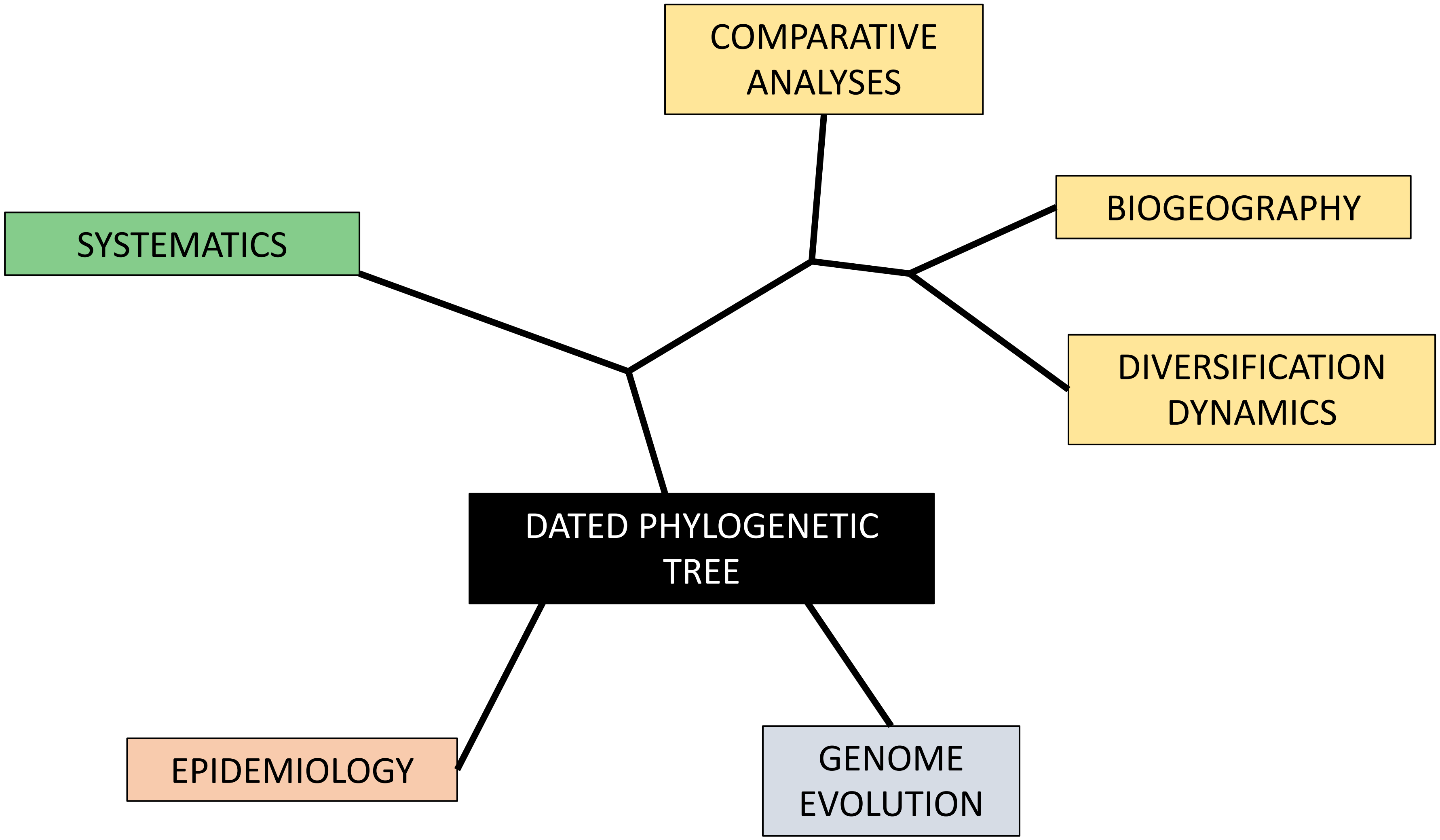
Bayesian methods include
a prior on tree growth

Branching process prior

Pure birth

Birth-death

- Test prior sensitivity on the results!
- Use model selection if feasible!



COMPARATIVE ANALYSES

BIOGEOGRAPHY

SYSTEMATICS

DIVERSIFICATION DYNAMICS

DATED PHYLOGENETIC TREE

EPIDEMIOLOGY

GENOME EVOLUTION

High Performance Computing at UH

- High Performance Computing at UH:
  - https://research.csc.fi/csc-s-servers
- Puhti supercomputer:
  - https://docs.csc.fi/computing/systems-puhti/

Bring your own computer to practical exercises next week!

Summary

- Ultimate aim: the combined-evidence analysis
- Bayesian Inference is a natural framework to incorporate fossils and date phylogeny
- Can be sensitive to the choice of priors and potential errors
- Manifold of various methods and software
- Tree dating allows addressing various biological questions