Workshop on Computational Genomics

Sun Yat-sen University Institute of Advanced Studies Hong Kong

March 4th 2025

Molecular Clock Dating

Dr Sandra Álvarez-Carretero

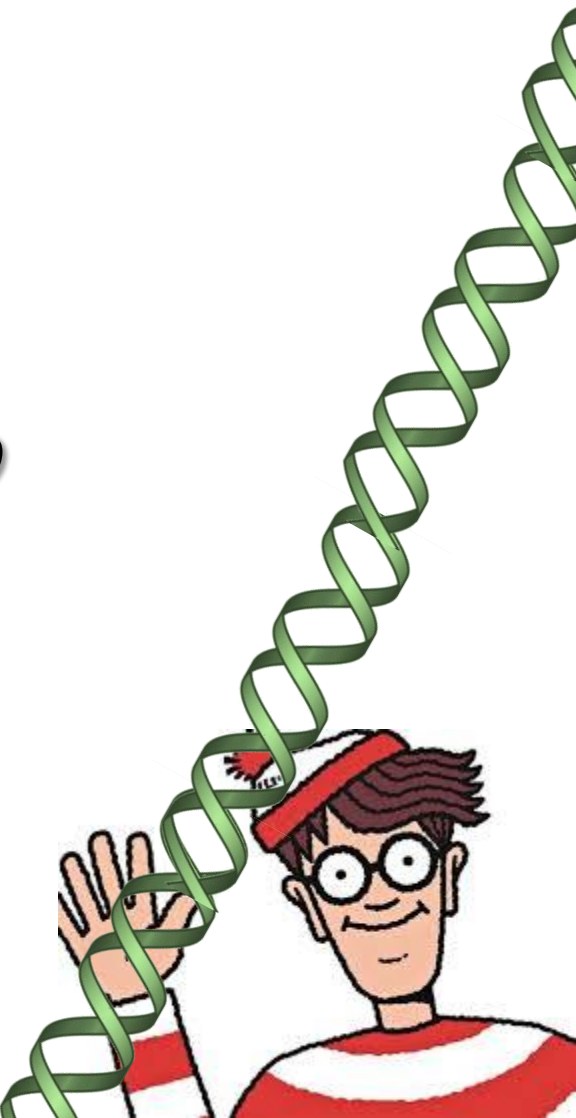
@sabifo4

What will we be covering during this session?

- Parameters to be estimated during timetree inference.
- Building intuition to set up the time prior using fossil/geological evidence.
- Building intuition to set up the rate prior.
- Why do we care about evolutionary timelines?
- Software for timetree inference.
- Approximating the likelihood calculation with `MCMCtree`.

WHERE ARE WE

in the phylogenetics workflow?



DOWNLOAD ORTHOLOGS FROM DATABASES

1. SEQUENCE SAMPLES AND OBTAIN RAW READS

Raw reads

2. FILTER RAW READS AND QUALITY CONTROL (QC)

Filtered reads

3. ASSEMBLE FILTERED READS AND QC

Genome assembly

4. GENE ANNOTATION

Gene A Gene B Gene C Gene D

5. INFER AND KEEP ONLY ORTHOLOGOUS GENES

Orthologs

SEQUENCE ALIGNMENT

Apply preferred partitioning scheme if needed

MODEL SELECTION

How does the evolutionary rate evolve?

Autocorrelated rate evolution

Independent rate evolution

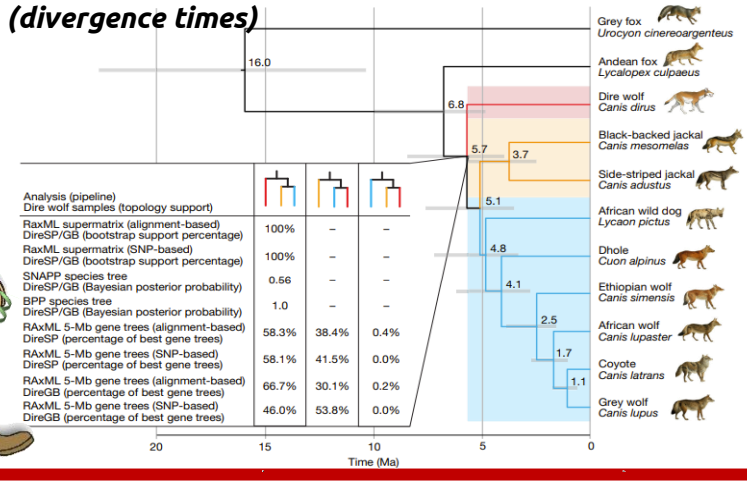
What substitution model do we need to use?

Some examples for nucleotide data:

JC69 K80 HKY85 ...

PHYLOGENY RECONSTRUCTION

INFERRING EVOLUTIONARY TIMELINE *(divergence times)*



Analysis (pipeline) Dire wolf samples (topology support)			
RaxML supermatrix (alignment-based) DireSP/GB (bootstrap support percentage)	100%	–	–
RaxML supermatrix (SNP-based) DireSP/GB (bootstrap support percentage)	100%	–	–
SNAPP species tree DireSP/GB (Bayesian posterior probability)	0.56	–	–
BPP species tree DireSP/GB (Bayesian posterior probability)	1.0	–	–
RAxML 5-Mb gene trees (alignment-based) DireSP (percentage of best gene trees)	58.3%	38.4%	0.4%
RAxML 5-Mb gene trees (SNP-based) DireSP (percentage of best gene trees)	58.1%	41.5%	0.0%
RAxML 5-Mb gene trees (alignment-based) DireGB (percentage of best gene trees)	66.7%	30.1%	0.2%
RAxML 5-Mb gene trees (SNP-based) DireGB (percentage of best gene trees)	46.0%	53.8%	0.0%

Revisiting
The Bayes' Theorem

with a focus on timetree inference

The Bayes' Theorem in timetree inference

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

$$P(param|data) = \frac{P(param)P(data|param)}{P(data)}$$

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

$$posterior = \frac{prior \times likelihood}{(marginal\ likelihood)}$$

The Bayes' Theorem in timetree inference

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

$$P(param|data) = \frac{P(param)P(data|param)}{P(data)}$$

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

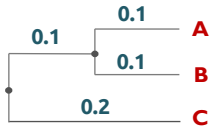
$$posterior = \frac{prior \times likelihood}{(marginal\ likelihood)}$$

What could θ be?

E.g., divergence times, evolutionary rate, tree topology, etc.

$\theta =$ 

Unknown parameters



If available, fossils can be informative about times!

If available, phylogenies (topology+branch lengths) can be informative about the rate!

The Bayes' Theorem in timetree inference

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

$$P(param|data) = \frac{P(param)P(data|param)}{P(data)}$$

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

$$posterior = \frac{prior \times likelihood}{(marginal\ likelihood)}$$

What could θ be?

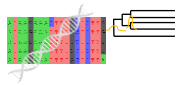
E.g., divergence times, evolutionary rate, tree topology, etc.

$\theta =$  **Unknown parameters**

Prior information on θ

What could D be?

E.g., molecular alignment, fixed tree topology, etc.

$D =$  **Data**

The Bayes' Theorem in timetree inference

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

$$P(param|data) = \frac{P(param)P(data|param)}{P(data)}$$

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

$$posterior = \frac{prior \times likelihood}{(marginal\ likelihood)}$$

What could θ be?

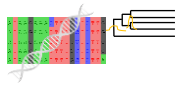
E.g., divergence times, evolutionary rate, etc.

$\theta =$  Unknown parameters

Prior information on θ

What could D be?

E.g., molecular alignment, fixed tree topology, etc.

$D =$  Data

Fossils can be data in **tip-dating analyses**, but we will not cover tip dating due to time limitations

The Bayes' Theorem in timetree inference

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

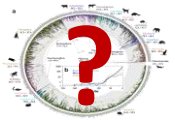
$$P(param|data) = \frac{P(param)P(data|param)}{P(data)}$$

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

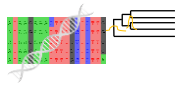
$$posterior = \frac{prior \times likelihood}{(marginal\ likelihood)}$$

Prior information on θ

EXAMPLE:

$\theta =$ 

Unknown parameters. E.g. in node–dating analyses: divergence times, evolutionary rate

$D =$ 

Data. E.g., in node–dating analyses: molecular data, fixed tree topology

$$f(\text{?}\,|\,\text{DNA}) = \frac{f(\text{?})\,f(\text{DNA}\,|\,\text{?})}{f(\text{DNA})}$$

Bayesian statistics applied to timetree inference analyses

CONDITIONAL PROBABILITY IN BAYES' THEOREM

$$posterior = \frac{prior \times likelihood}{(marginal\ likelihood)}$$

$$f(t, r, \theta | D) = \frac{f(\theta) f(t) f(r \mid t, \theta)\ f(D|t, r, \theta)}{f(D)}$$

D = molecular data
t = vector of divergence times
r = vector of molecular rates
θ = vector of other unknown parameter/s

$$f(\text{?} \mid \text{DNA}) = \frac{f(\text{?})\ f(\text{DNA} \mid \text{?})}{f(\text{DNA})}$$

Bayesian statistics applied to timetree inference analyses

CONDITIONAL PROBABILITY IN BAYES' THEOREM

$$posterior = \frac{prior \times likelihood}{(marginal\ likelihood)}$$

$$f(t, r, \theta|D) = \frac{f(\theta)f(t)f(r \mid t, \theta)\, f(D|t, r, \theta)}{f(D)}$$

$$f(t, r|D) = \frac{f(t)f(r \mid t)\, f(D|t, r)}{\iint f(t)f(r \mid t)\, f(D|t, r)\, \mathbf{dr}\, \mathbf{dt}}$$

D = molecular data
$\boldsymbol{t}$ = vector of divergence times
$\boldsymbol{r}$ = vector of molecular rates
$\boldsymbol{\theta}$ = vector of other unknown parameter/s

Bayesian statistics applied to timetree inference analyses

CONDITIONAL PROBABILITY IN BAYES' THEOREM

$$posterior = \frac{prior \times likelihood}{(marginal\ likelihood)}$$

$$f(\boldsymbol{t}, \boldsymbol{r}, \boldsymbol{\theta}|D) = \frac{f(\boldsymbol{\theta}) f(\boldsymbol{t}) f(\boldsymbol{r} \mid \boldsymbol{t}, \boldsymbol{\theta})\ f(D|\boldsymbol{t}, \boldsymbol{r}, \boldsymbol{\theta})}{f(D)}$$

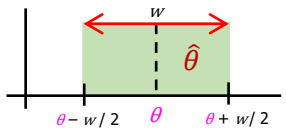
$$f(\boldsymbol{t}, \boldsymbol{r}|D) = \frac{f(\boldsymbol{t}) f(\boldsymbol{r} \mid \boldsymbol{t})\ f(D|\boldsymbol{t}, \boldsymbol{r})}{\iint f(\boldsymbol{t}) f(\boldsymbol{r} \mid \boldsymbol{t})\ f(D|\boldsymbol{t}, \boldsymbol{r})\ \mathbf{dr}\ \mathbf{dt}}$$

D = molecular data
$\boldsymbol{t}$ = vector of divergence times
$\boldsymbol{r}$ = vector of molecular rates
$\boldsymbol{\theta}$ = vector of other unknown parameter/s

Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo – *how does it work?*

0. Set initial value for model parameters to be estimated. In addition, specify number of iterations, n, and create vector `theta` to collect sampled values.
1. Calculate prior, $f(\theta)$.
2. Calculate likelihood, $f(x|\theta)$.
3. Calculate unnormalised posterior (i.e., $f(\theta|x) = f(\theta) \times f(x|\theta)$).
4. Proposal: sample a random parameter value under a uniform distribution (or another) to get the new proposal $\hat{\theta}$. If $\hat{\theta} < 0$, then $\hat{\theta} = -\hat{\theta}$.



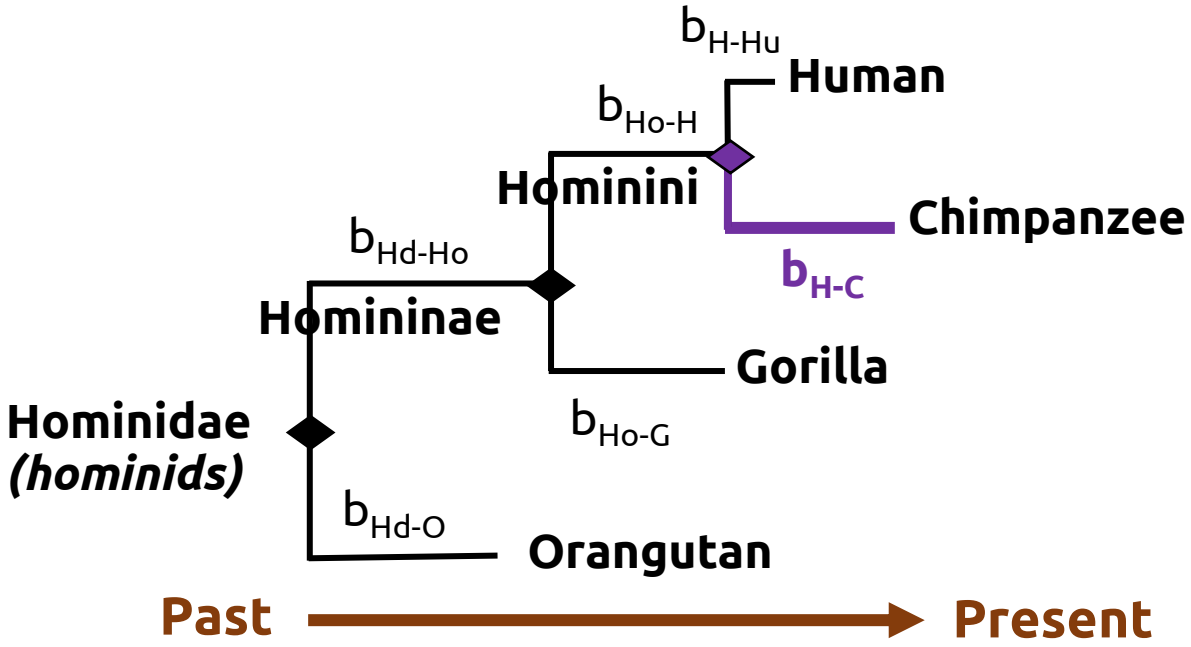
5. Calculate prior', $f(\hat{\theta})$, likelihood', $f(x|\hat{\theta})$, and unnormalised posterior' with new proposed value $\hat{\theta}$, $f(\hat{\theta}|x)$.
6. Accept or reject $\hat{\theta}$ value. If accepted, $\theta \leftarrow \hat{\theta}$. Otherwise, keep initial value for the next iteration $\theta \leftarrow \theta$.
7. Save value of θ in vector `theta`.
8. Repeat 1-7 n times with final θ according to step 7.
9. Return vector `theta` with sampled θ values. Plot traces, histograms, etc. to assess chain mixing, efficiency, and convergence).

Our parameters of interest θ are now divergence times, t, and evolutionary rates, r !

SO…

How can we estimate rates and times?

Understanding the molecular clock



branch length = evolutionary rate x divergence time

$$b_{H-C} = r_{H-C} \times t_{H-C}$$

Branch lengths are like a "clock": they help us understand when and at which rate evolution has taken place

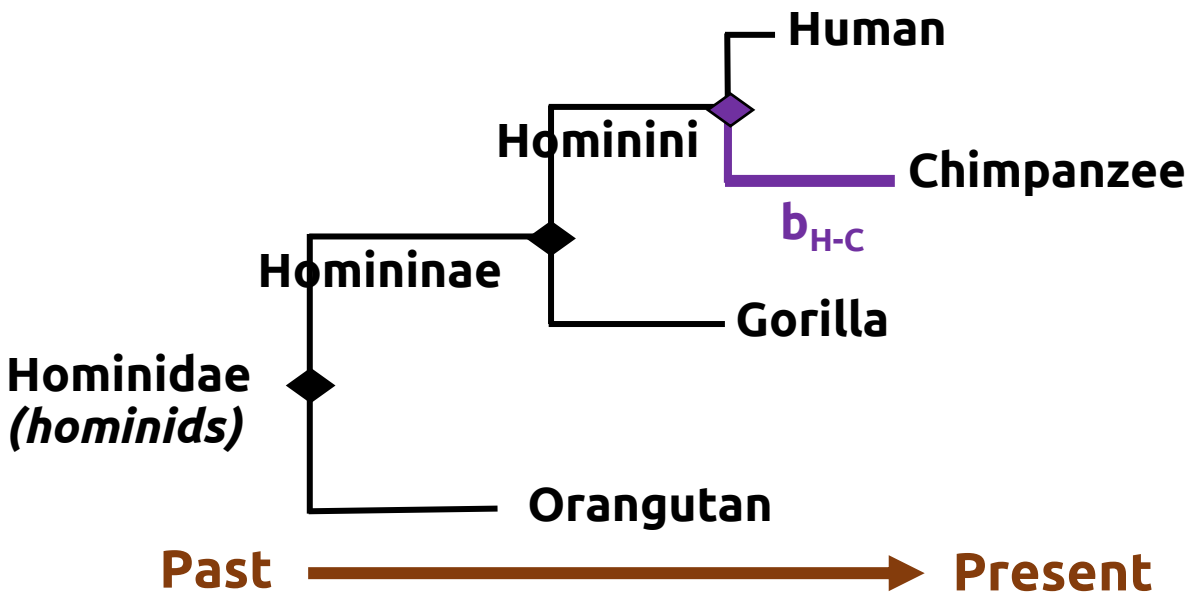
Divergence times *(t)*

When two species are biologically distinct, they have diverged

Evolutionary rate *(r)*

How often do mutations accumulate through time?

Understanding the molecular clock



branch length = evolutionary rate x divergence time

$$b_{H-C} = r_{H-C} \times t_{H-C}$$

Branch length Hominini-Chimpanzee

$b_{H-C} = r_{H-C} \times t_{H-C}$

Let's imagine there are 6 mutations per site per time unit:

$6 = r_{H-C} \times t_{H-C}$

$6 = 1 \times 6$

$6 = 6 \times 1$

$6 = 2 \times 3$

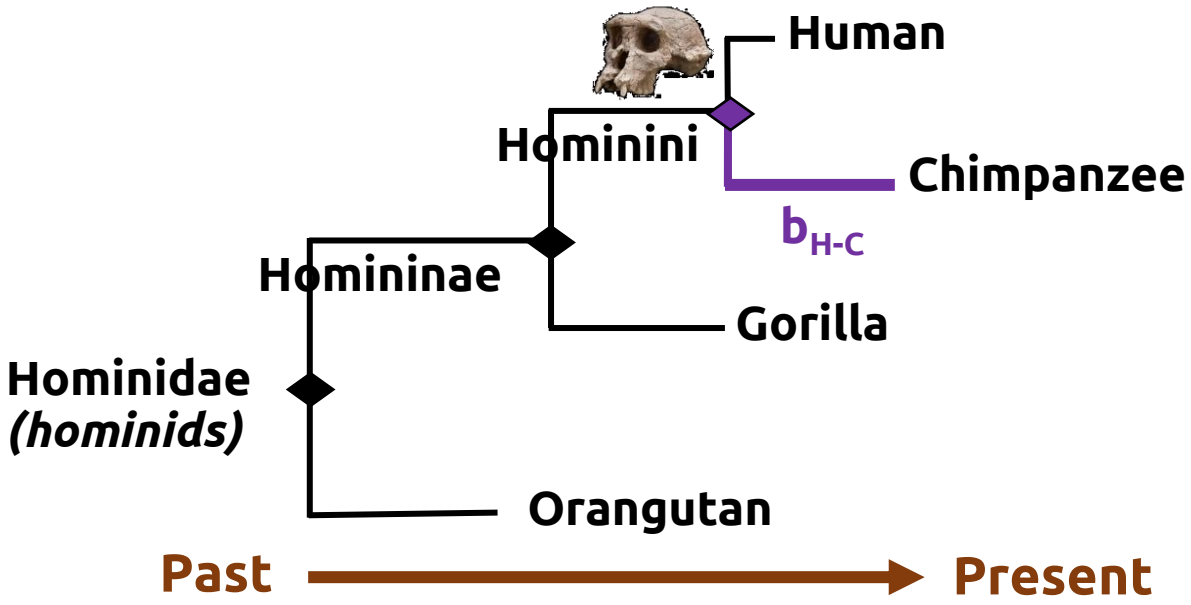
…

More than one plausible solution…

We need additional info to estimate rates and times separately!

PROBLEM: current methods estimate branch lengths, and so times and rates are confounded!

Calibrating the molecular clock



branch length = evolutionary rate x divergence time

$$b_{H-C} = r_{H-C} \times t_{H-C}$$

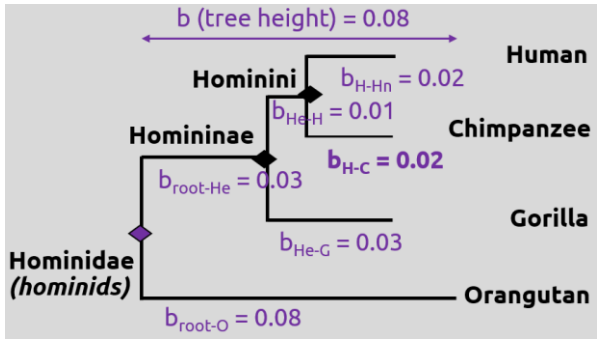
Thanks to *a priori* information, we can integrate the uncertainty about estimates of divergence times, evolutionary rates, and branch lengths through the usage of PRIORS

FOSSIL/GEOLOGICAL EVIDENCE CAN BE USEFUL!

EXAMPLE

- †*Sahelanthropus*, common ancestor of chimpanzee and human
- Minimum age: 5.333 Ma
- Maximum age: 7.246 Ma
- If more than one specimen, we use the oldest!

PHYLOGENIES CAN BE USEFUL!

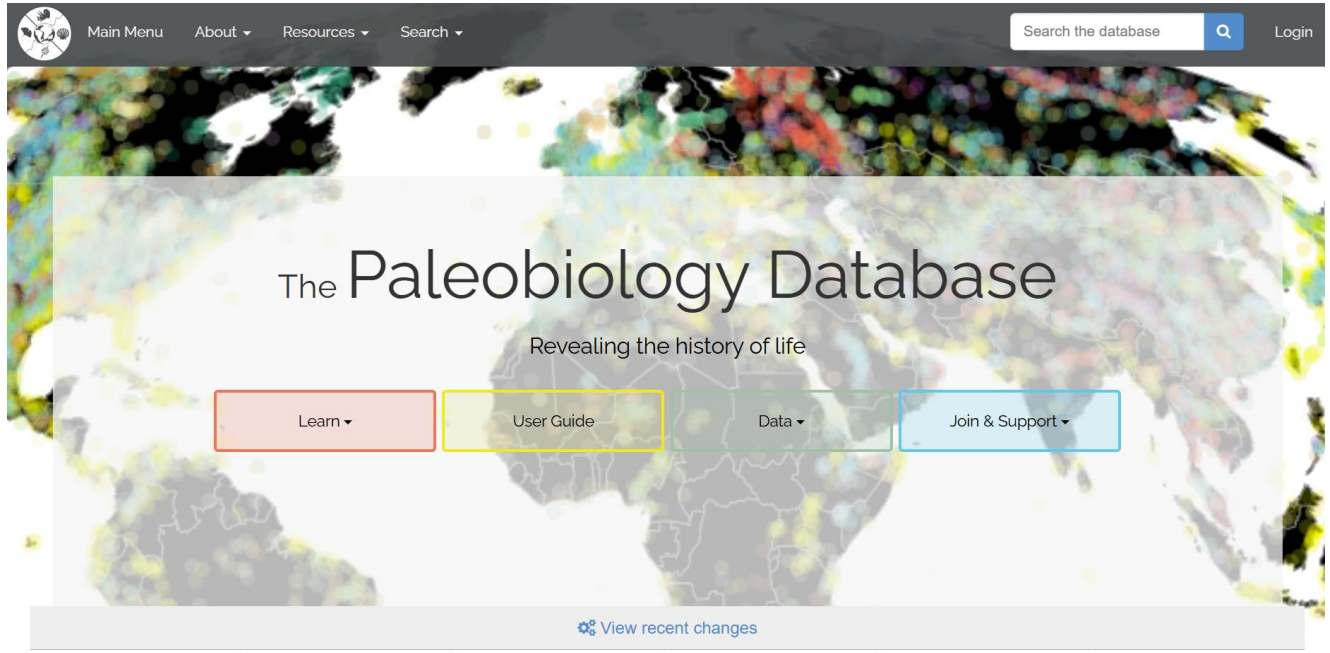


We can estimate the evolutionary rate or use other people's estimated values!

TIME PRIOR

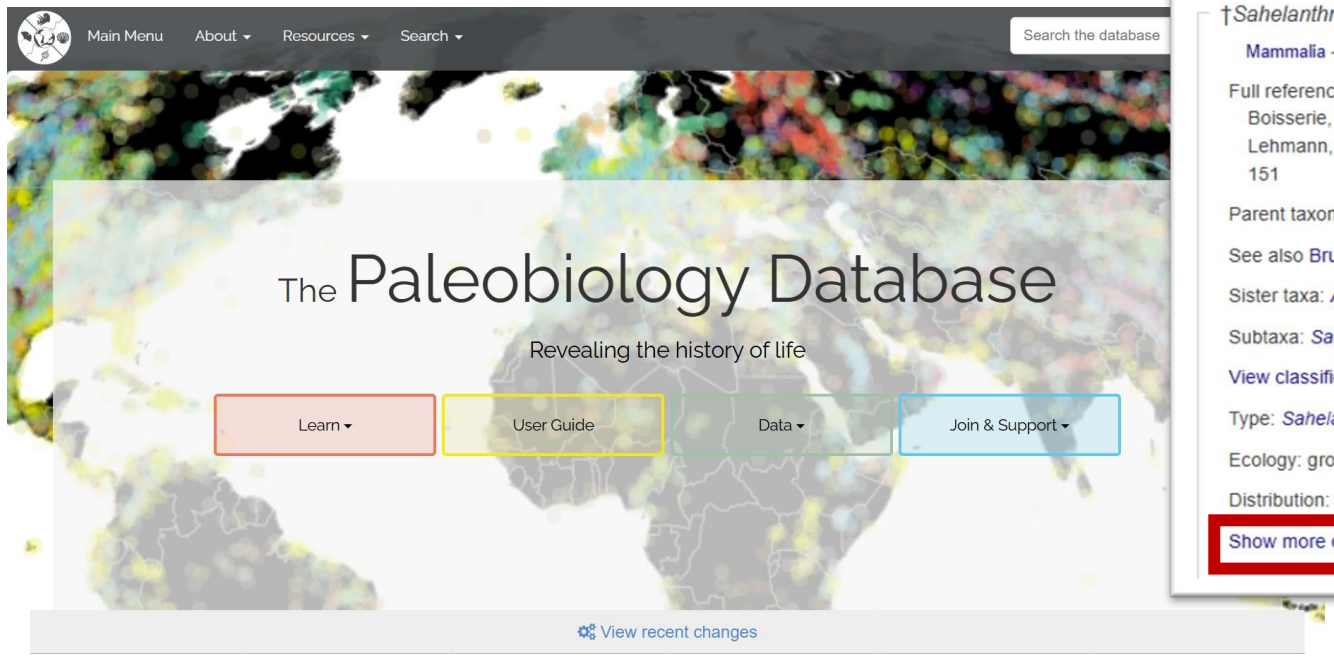
Where can we get fossil information?

- **Search the literature**. E.g., relevant papers published about the fossil specimen you want to incorporate in your study
- **Collaborate with experts (palaeontologists, geologists, etc.)**
- **Use the Paleobiology Database (PBDB)**: this is the main database that you can use to track the many fossil specimens that have been discovered and catalogued and is the main site to store fossil information! URL: https://paleobiodb.org/



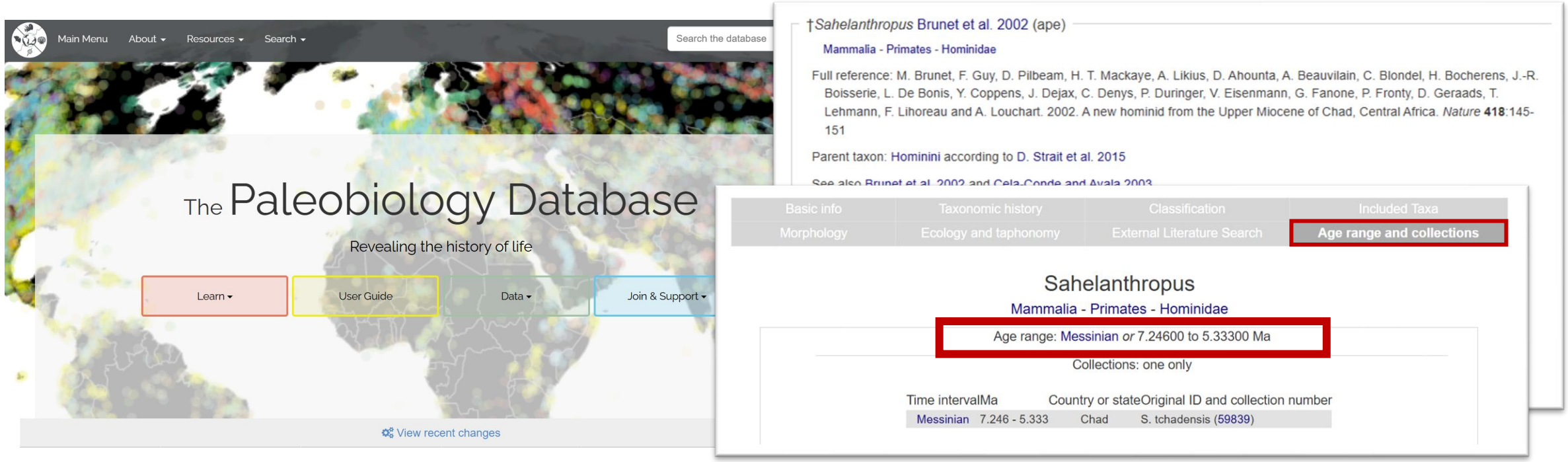
Where can we get fossil information?

- **Search the literature**. E.g., relevant papers published about the fossil specimen you want to incorporate in your study
- **Collaborate with experts (palaeontologists, geologists, etc.)**
- **Use the Paleobiology Database (PBDB)**: this is the main database that you can use to track the many fossil specimens that have been discovered and catalogued and is the main site to store fossil information! URL: https://paleobiodb.org/

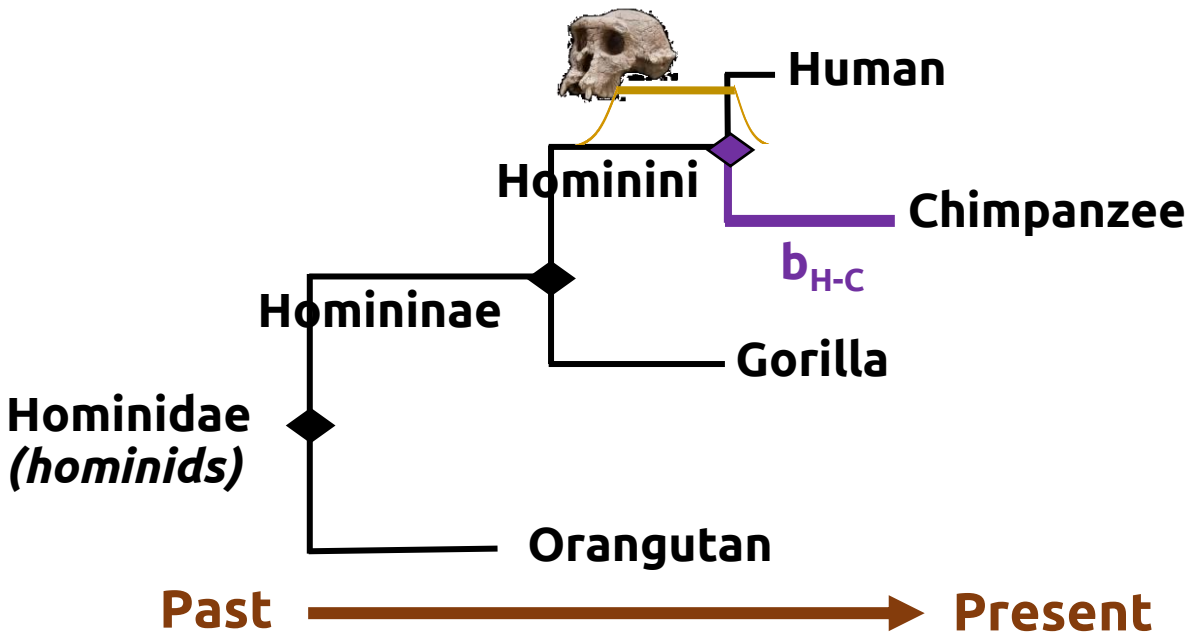


Where can we get fossil information?

- **Search the literature**. E.g., relevant papers published about the fossil specimen you want to incorporate in your study
- **Collaborate with experts (palaeontologists, geologists, etc.)**
- **Use the Paleobiology Database (PBDB)**: this is the main database that you can use to track the many fossil specimens that have been discovered and catalogued and is the main site to store fossil information! URL: https://paleobiodb.org/

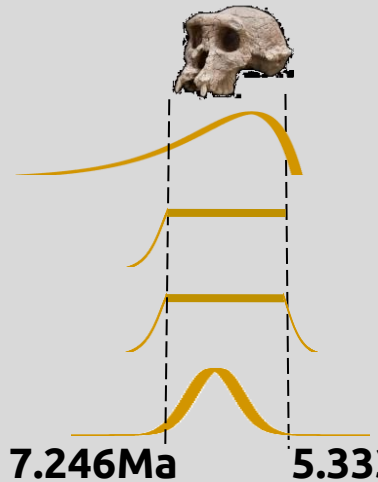


Setting the time prior



branch length = evolutionary rate x divergence time

$$b_{H-C} = r_{H-C} \times t_{H-C}$$

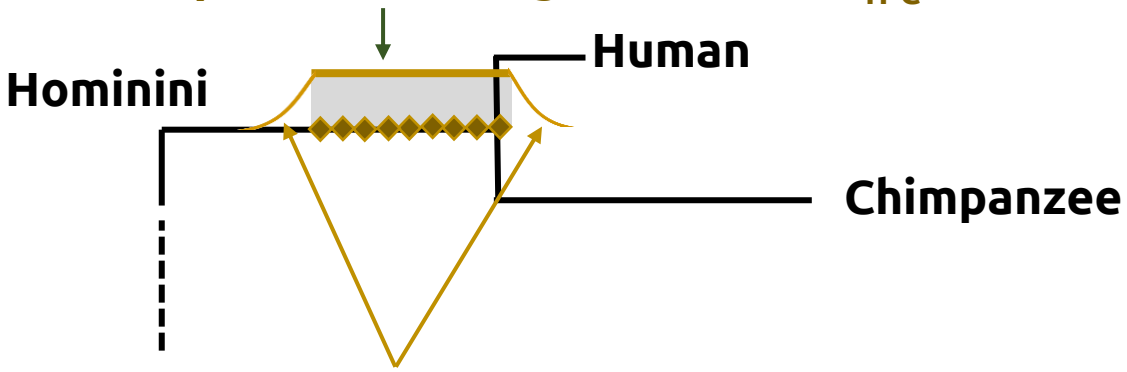


E.g.: possible "priors" to integrate the uncertainty about the fossil record

Thanks to *a priori* information, we can integrate the uncertainty about estimates of divergence times, evolutionary rates, and branch lengths through the usage of PRIORS

Setting the time prior

More plausible divergence times, $t_{H\text{-}C}$



Less plausible divergence times, $t_{H\text{-}C}$
Probability decreases with a tail percentage of ~2.5%!

branch length = evolutionary rate x divergence time

$$b_{H\text{-}C} = r_{H\text{-}C} \times t_{H\text{-}C}$$

$$posterior \propto priors \times likelihood$$

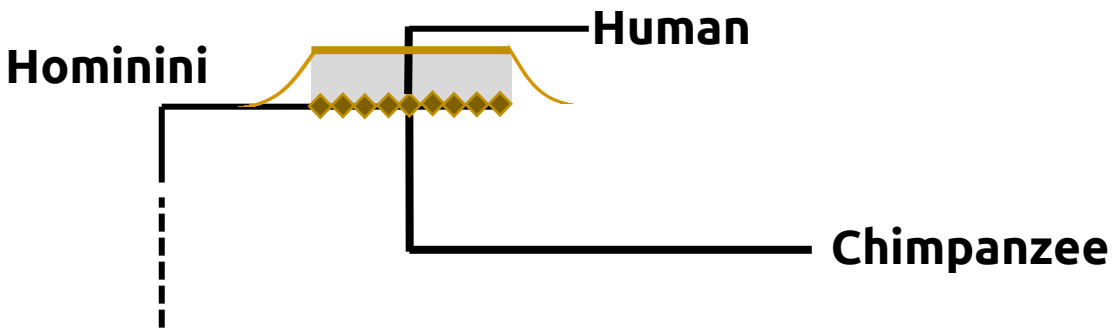
$$f(t, r|D) \propto f(t) \times f(r|t) \times f(D|t, r)$$

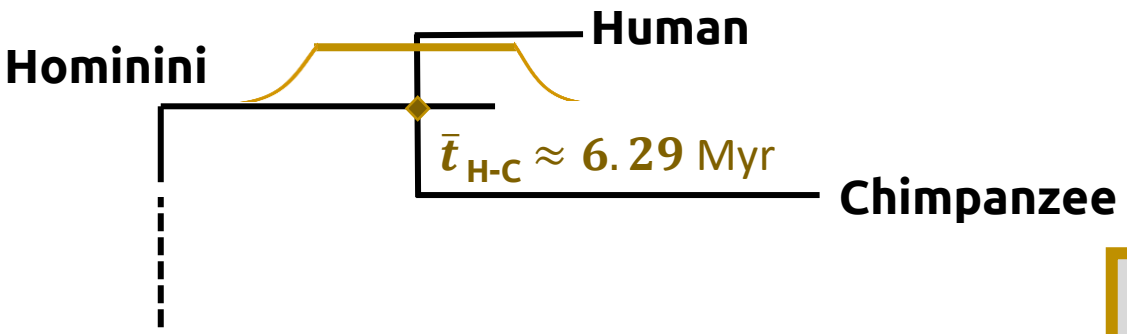
TIME PRIOR (t)

7.246Ma 5.333 Ma

Uncertainty in the fossil record to estimate t

Setting the time prior





By using ONLY our knowledge on the fossil record, our mean estimated divergence time is $\bar{t}_{H\text{-}C}$ = 6.29 Myr before we add molecular data in the analysis

branch length = evolutionary rate x divergence time

$$b_{H\text{-}C} = r_{H\text{-}C} \times t_{H\text{-}C}$$

$$posterior \propto priors \times likelihood$$

$$f(t, r|D) \propto f(t) \times f(r|t) \times f(D|t, r)$$

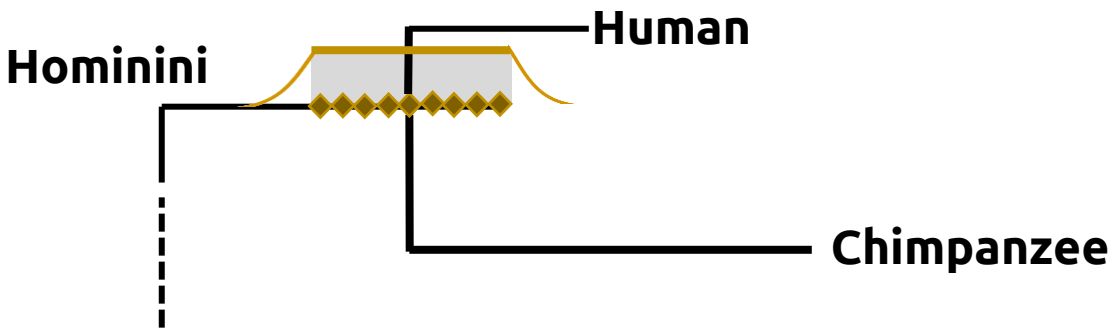
TIME PRIOR (t)

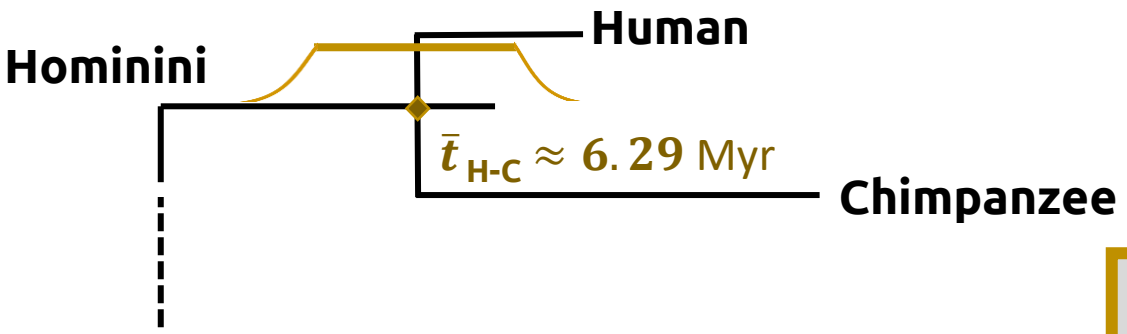
$\bar{t} \approx 6.29$ Myr

7.246Ma 5.333 Ma

Uncertainty in the fossil record to estimate t

Setting the time prior





By using ONLY our knowledge on the fossil record, our mean estimated divergence time is $\bar{t}_{H\text{-}C}$ = 6.29 Myr before we add molecular data in the analysis

branch length = evolutionary rate x divergence time

$$b_{H\text{-}C} = r_{H\text{-}C} \times t_{H\text{-}C}$$

$$posterior \propto priors \times likelihood$$

$$f(t, r|D) \propto f(t) \times f(r|t) \times f(D|t, r)$$

TIME PRIOR (t)

$\bar{t} \approx 6.29$ Myr

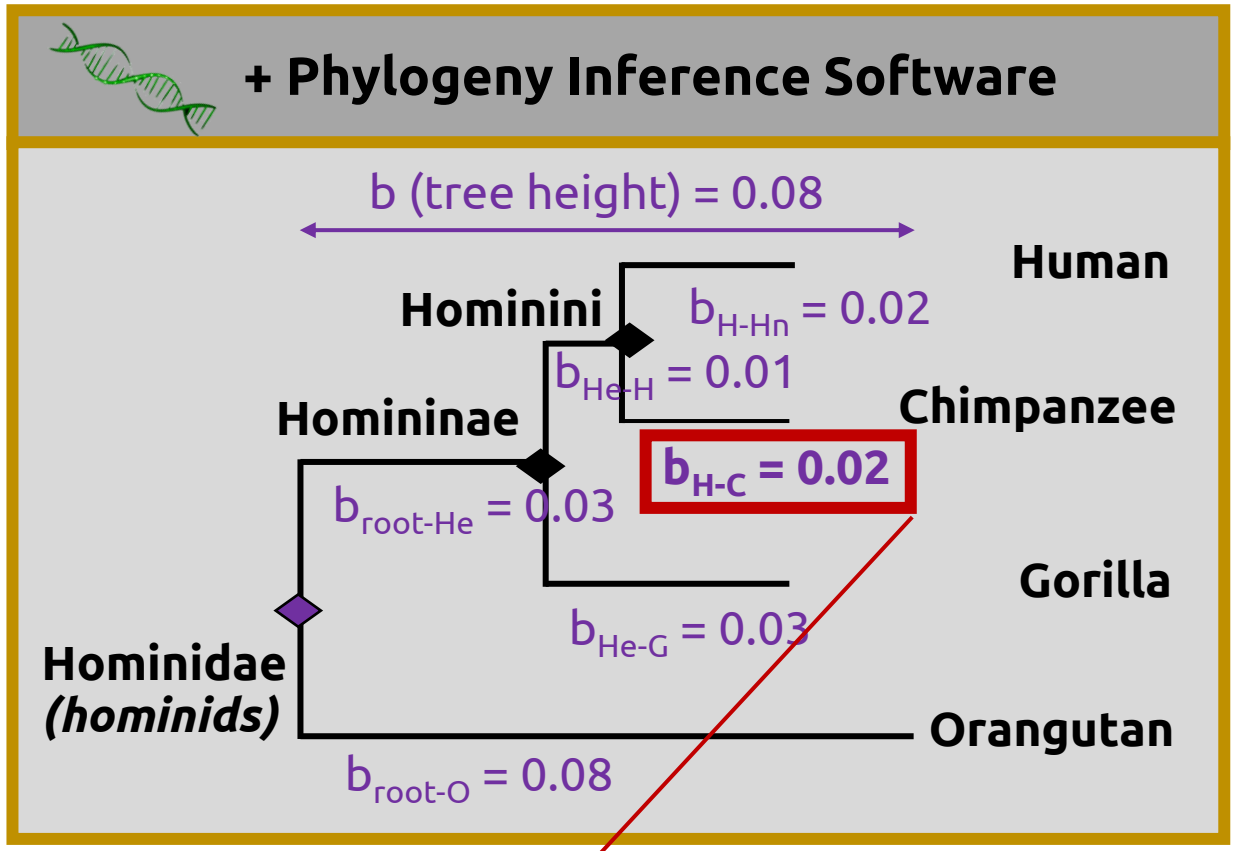
7.246Ma 5.333 Ma

Uncertainty in the fossil record to estimate t

NOTE: this is just an example to build intuition on how you could gain some prior information on the divergence times of a node without using sequence data; the time prior we use in timetree inference is actually more complicated than that!

RATE PRIOR

Setting the rate prior



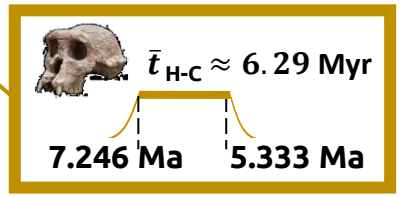
branch length = evolutionary rate x divergence time

$b_{H\text{-}C} = r_{H\text{-}C} \times t_{H\text{-}C}$

$b_{H\text{-}C} = r_{H\text{-}C} \times t_{H\text{-}C}$

$r_{H\text{-}C} = b_{H\text{-}C} / t_{H\text{-}C}$ = **0.02** *subst/site* **/ 8.25** *Myr*

$r_{H\text{-}C}$ = **0.002** *subst/site/Myr*



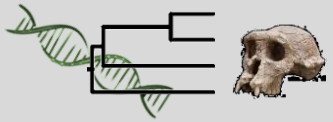
$$posterior \propto priors \times likelihood$$

$$f(t, r|D) \propto f(t) \times f(r|t) \times f(D|t, r)$$

By combining our prior on times ($\bar{t}_{H\text{-}C}$) and our recently gained knowledge on the branch lengths ($\hat{b}_{H\text{-}C}$), our estimated mean rate before we include molecular data in the analysis is $\bar{r}_{H\text{-}C}$ = 0.002 s/s/Myr

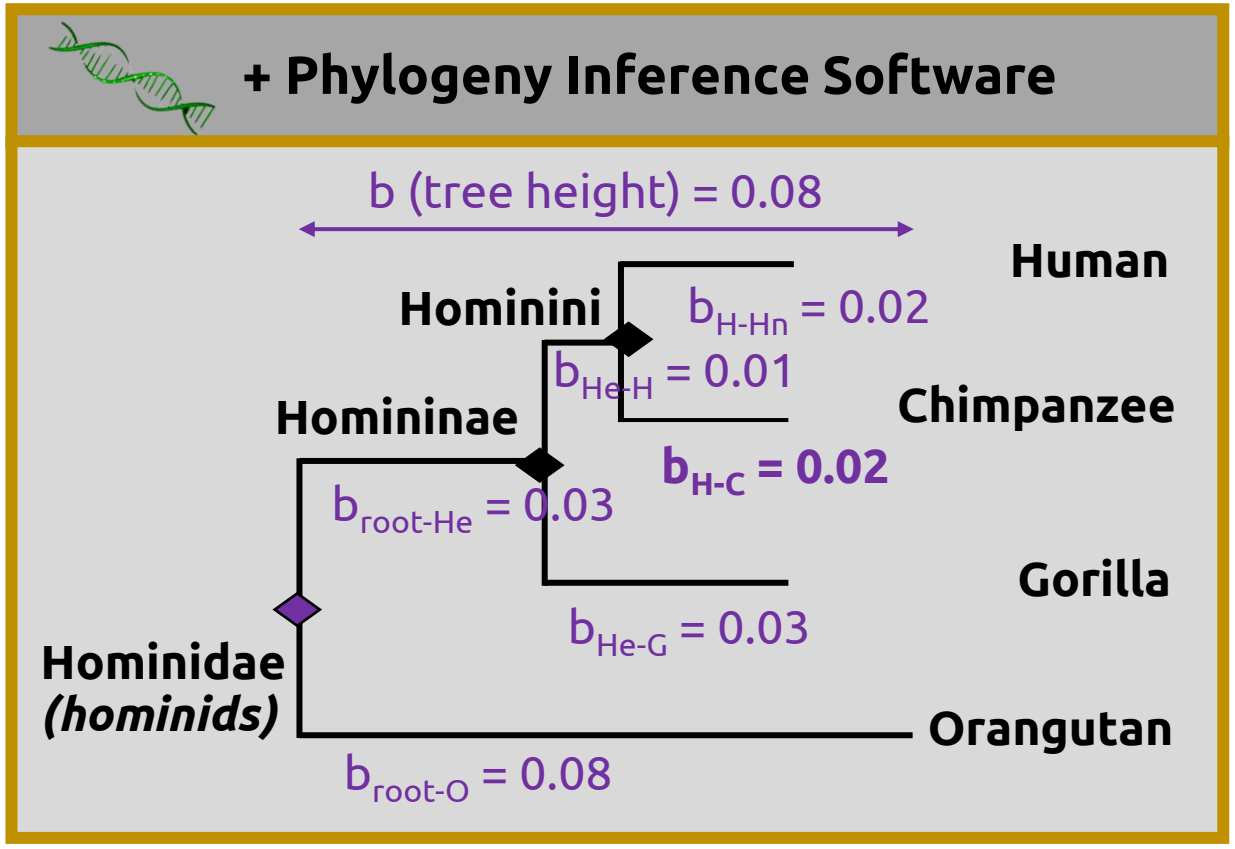
RATE PRIOR (r)

$\bar{r} \approx$ 0.002 s/s/t



Uncertainty in e.g. data, branch lengths, etc. used to estimate r

Setting the rate prior



The clock only holds for closely-related species, otherwise, it is violated -- not a good hypothesis!

Current approaches use relaxed-clock models to allow for the fact that species in a phylogeny may evolve at different rates!

branch length = evolutionary rate x divergence time

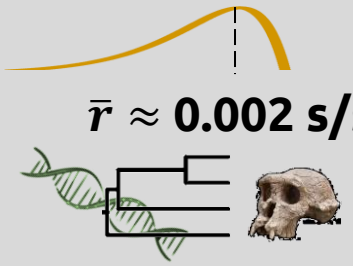
$$b_{H\text{-}C} = r_{H\text{-}C} \times t_{H\text{-}C}$$

$$posterior \propto priors \times likelihood$$

$$f(t, r|D) \propto f(t) \times f(r|t) \times f(D|t, r)$$

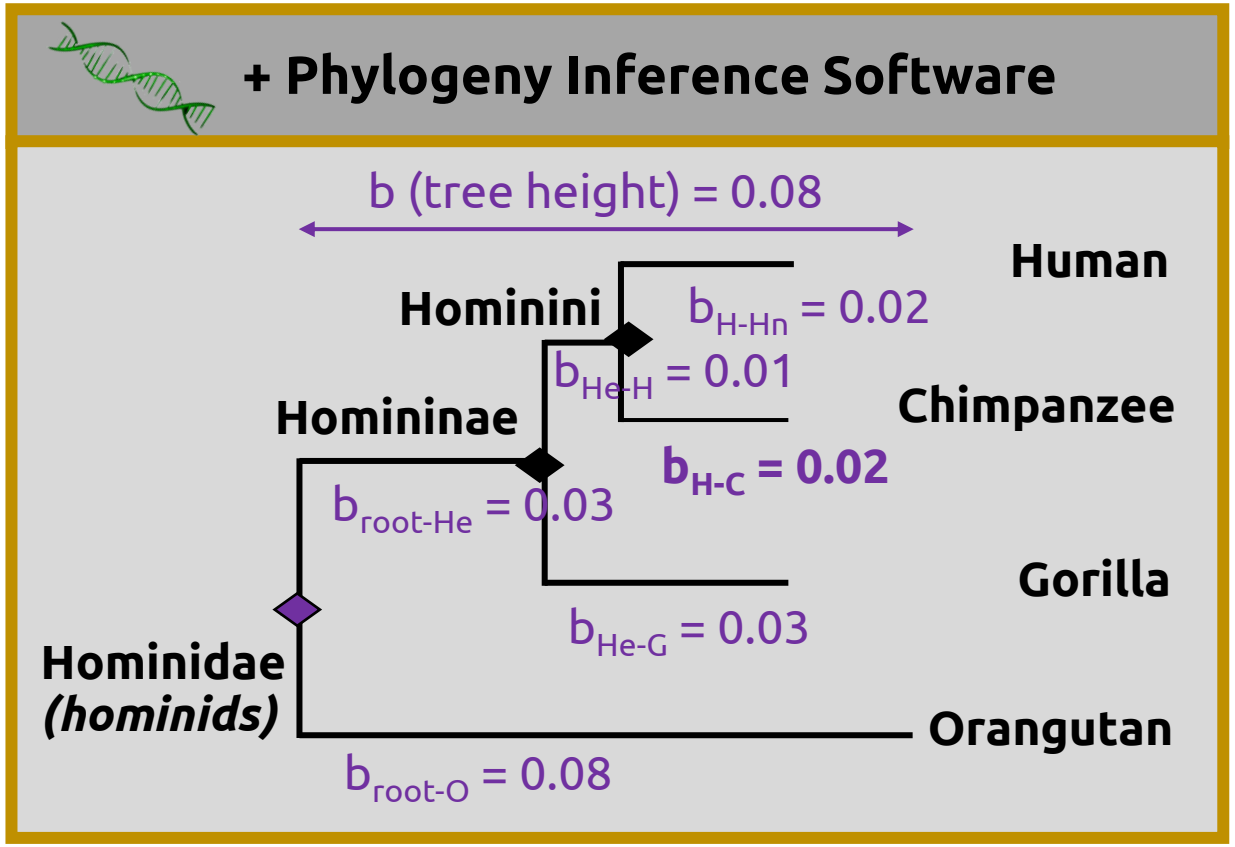
By combining our prior on times ($\bar{t}_{H\text{-}C}$) and our recently gained knowledge on the branch lengths ($\hat{b}_{H\text{-}C}$), our estimated mean rate before we include molecular data in the analysis is $\bar{r}_{H\text{-}C}$ = 0.002 s/s/Myr

RATE PRIOR (r)



Uncertainty in e.g. data, branch lengths, etc. used to estimate r

Setting the rate prior



The clock only holds for closely-related species, otherwise, it is violated -- not a good hypothesis!

Current approaches use relaxed-clock models to allow for the fact that species in a phylogeny may evolve at different rates!

branch length = evolutionary rate x divergence time

$$b_{H\text{-}C} = r_{H\text{-}C} \times t_{H\text{-}C}$$

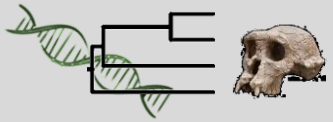
$$posterior \propto priors \times likelihood$$

$$f(t, r|D) \propto f(t) \times f(r|t) \times f(D|t, r)$$

NOTE: this is just an example to build intuition on how you could gain some prior information on a given branch rate; the rate prior we use in timetree inference is actually more complicated than that!

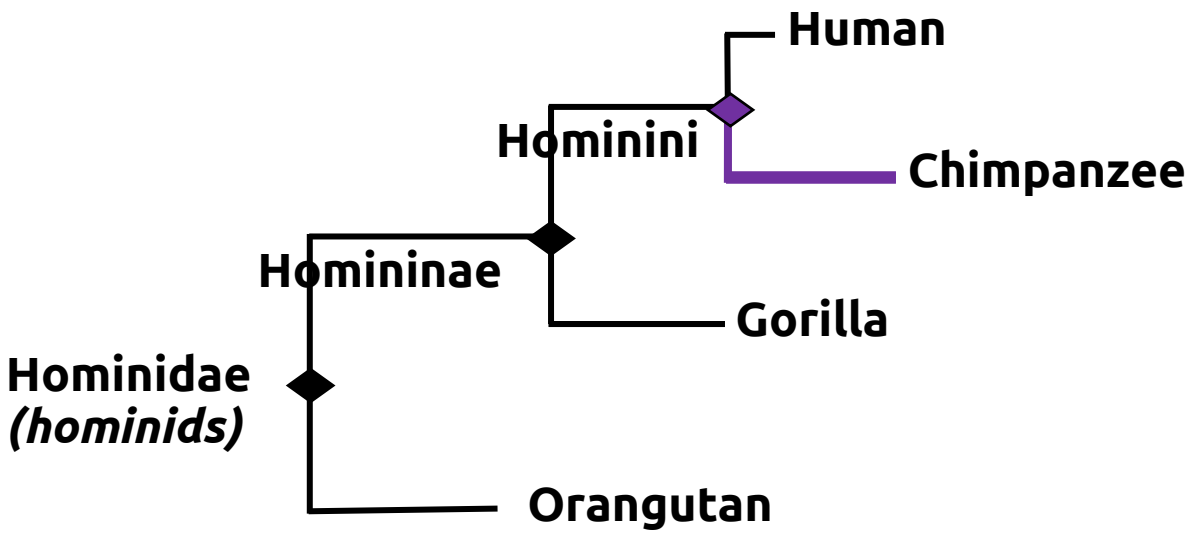
RATE PRIOR (r)

$\bar{r} \approx 0.002$ s/s/t



Uncertainty in e.g. data, branch lengths, etc. used to estimate r

Likelihood



branch length = evolutionary rate x divergence time

$b_{H\text{-}C} = r_{H\text{-}C} \times t_{H\text{-}C}$

$$posterior \propto priors \times likelihood$$

$$f(t, r|D) \propto f(t) \times f(r|t) \times f(D|t, r)$$

TIME PRIOR (t)

$\bar{t} \approx 6.29$ Myr

7.246Ma 5.333 Ma

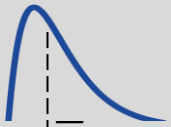
Uncertainty in the fossil record to estimate t

RATE PRIOR (r)

$\bar{r} \approx 0.002$ s/s/t

Uncertainty in e.g. data, branch lengths, etc. used to estimate r

BRANCH LENGTH UNCERTAINTY

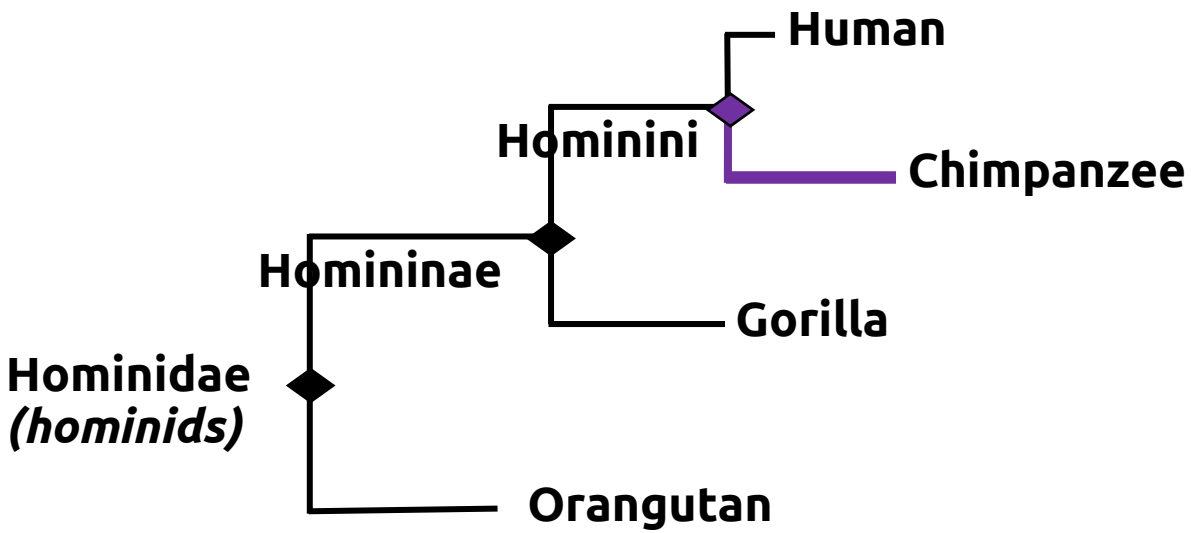


$\bar{b} \approx 0.02$ s/s

? $t_{H\text{-}C} = 8.25$ Myr
$r_{H\text{-}C} = 0.02$ s/s/Myr

Combined uncertainty in estimates of t and r (i.e., $b = r \times t$)

Likelihood



branch length = evolutionary rate x divergence time

$b_{H\text{-}C} = r_{H\text{-}C} \times t_{H\text{-}C}$

$$posterior \propto priors \times likelihood$$

$$f(t, r|D) \propto f(t) \times f(r|t) \times f(D|t, r)$$

NOTE: this is just an example to build intuition on how you could calculate the likelihood given your updated knowledge on rates and times; the likelihood function we use in timetree inference is actually more complicated than that!

TIME PRIOR (t)

$\bar{t} \approx 6.29$ Myr

7.246Ma 5.333 Ma

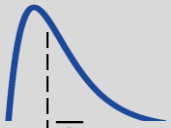
Uncertainty in the fossil record to estimate t

RATE PRIOR (r)

$\bar{r} \approx 0.002$ s/s/t

Uncertainty in e.g. data, branch lengths, etc. used to estimate r

BRANCH LENGTH UNCERTAINTY

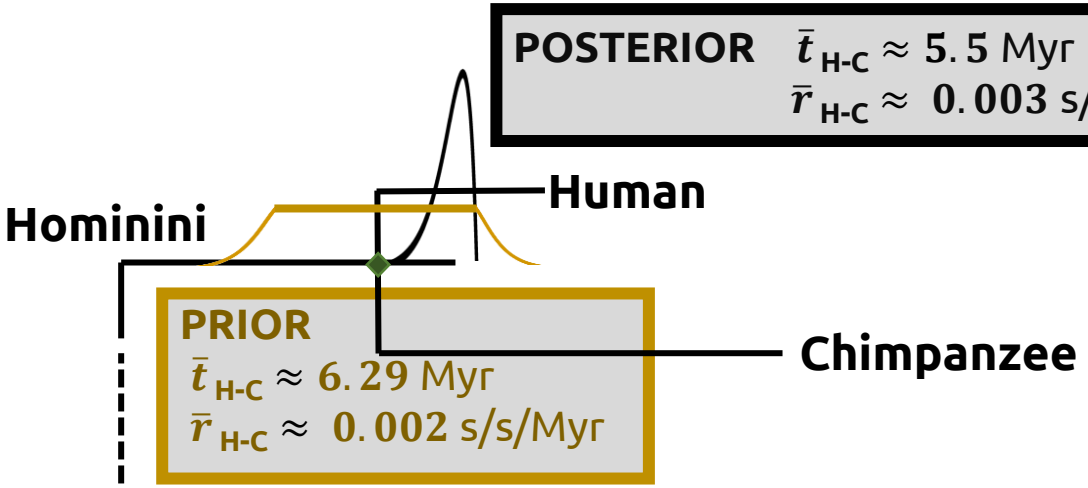


$\bar{b} \approx 0.02$ s/s

? $t_{H\text{-}C} = 8.25$ Myr
$r_{H\text{-}C} = 0.02$ s/s/Myr

Combined uncertainty in estimates of t and r (i.e., $b = r \times t$)

Estimating posterior densities (rates and times)



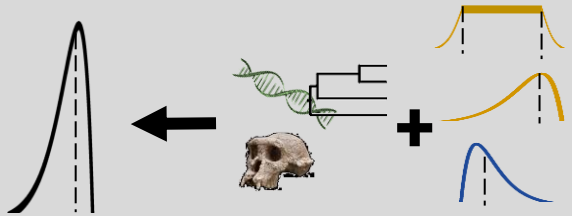
branch length = evolutionary rate x divergence time

$b_{H\text{-}C} = r_{H\text{-}C} \times t_{H\text{-}C}$

$$posterior \propto priors \times likelihood$$

$$f(t, r|D) \propto f(t) \times f(r|t) \times f(D|t, r)$$

Posterior (t, r)



Estimate mean posterior t and r, and corresponding CIs!

TIME PRIOR (t)

$\bar{t} \approx 6.29$ Myr

7.246Ma 5.333 Ma

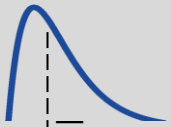
Uncertainty in the fossil record to estimate t

RATE PRIOR (r)

$\bar{r} \approx 0.002$ s/s/t

Uncertainty in e.g. data, branch lengths, etc. used to estimate r

BRANCH LENGTH UNCERTAINTY



$\bar{b} \approx 0.02$ s/s

? $t_{H\text{-}C} = 8.25$ Myr
$r_{H\text{-}C} = 0.02$ s/s/Myr

Combined uncertainty in estimates of t and r (i.e., $b = r \times t$)

Estimating posterior densities (rates and times)

NOTE: as mentioned before, this example has been used with the aim to build intuition on how you could gain prior information on the rates and the times for a specific lineage, calculate the likelihood given your updated knowledge on the rate and the divergence time, and, lastly, put everything together to estimate the posterior densities of our parameters of interest. Nevertheless, estimating these parameters in timetree inference is of course much more convoluted than that!

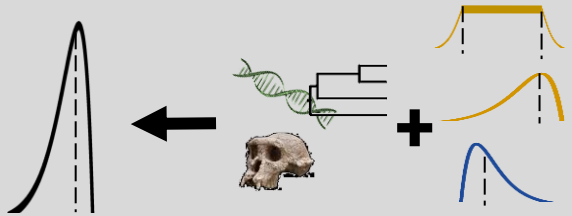
branch length = evolutionary rate x divergence time

$b_{H\text{-}C} = r_{H\text{-}C} \times t_{H\text{-}C}$

$$posterior \propto priors \times likelihood$$

$$f(t, r|D) \propto f(t) \times f(r|t) \times f(D|t, r)$$

Posterior (t, r)



Estimate mean posterior t and r, and corresponding CIs!

TIME PRIOR (t)

$\bar{t} \approx 6.29$ Myr

7.246Ma 5.333 Ma

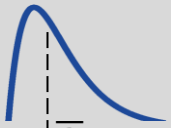
Uncertainty in the fossil record to estimate t

RATE PRIOR (r)

$\bar{r} \approx 0.002$ s/s/t

Uncertainty in e.g. data, branch lengths, etc. used to estimate r

BRANCH LENGTH UNCERTAINTY



$\bar{b} \approx 0.02$ s/s

? $t_{H\text{-}C} = 8.25$ Myr
$r_{H\text{-}C} = 0.02$ s/s/Myr

Combined uncertainty in estimates of t and r (i.e., $b = r \times t$)

BUT…

Why do we want to infer evolutionary timelines?

Are they useful at all?

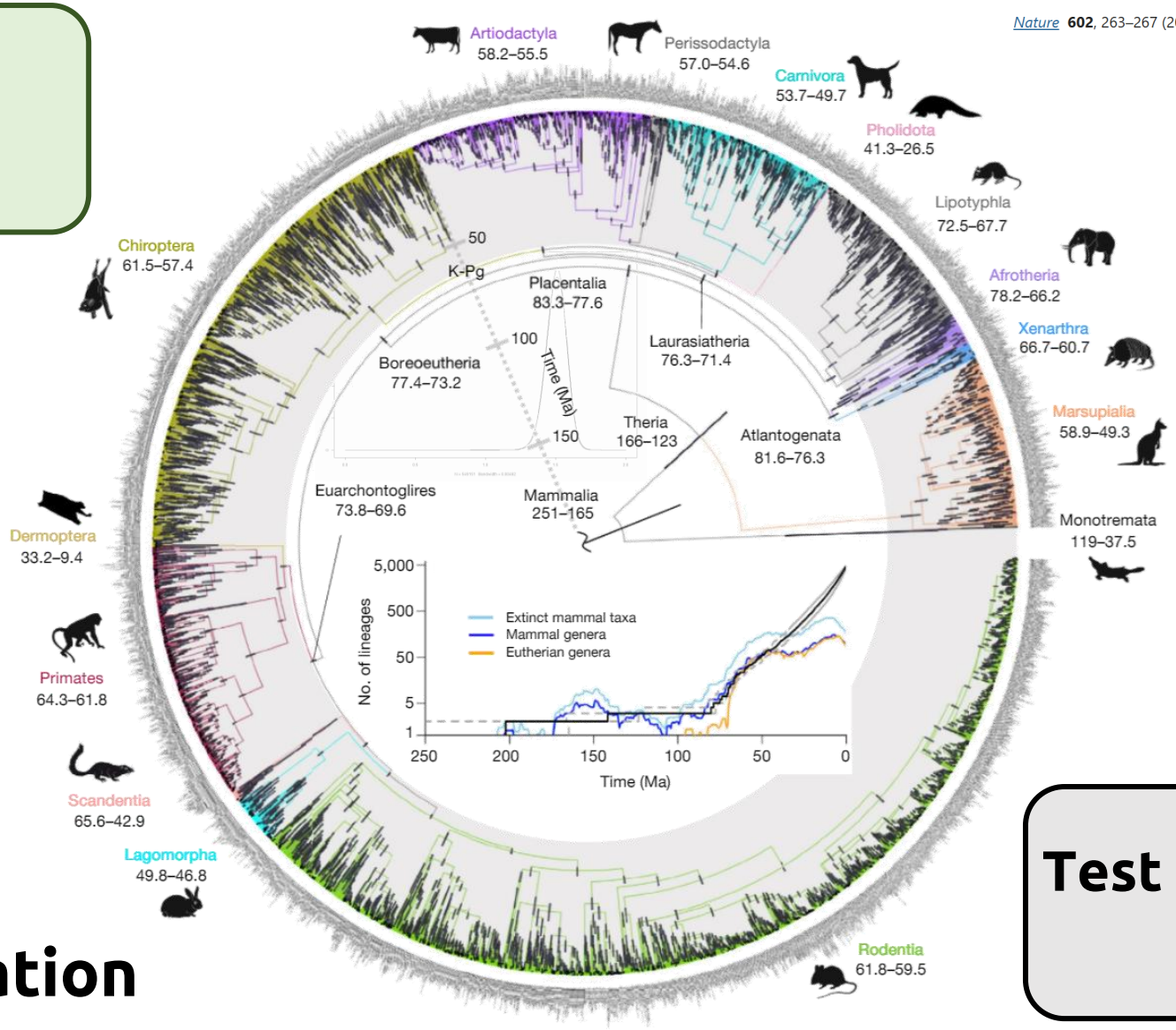
Why are evolutionary timelines useful?

Study biodiversity

Evolution

Ecology

Conservation



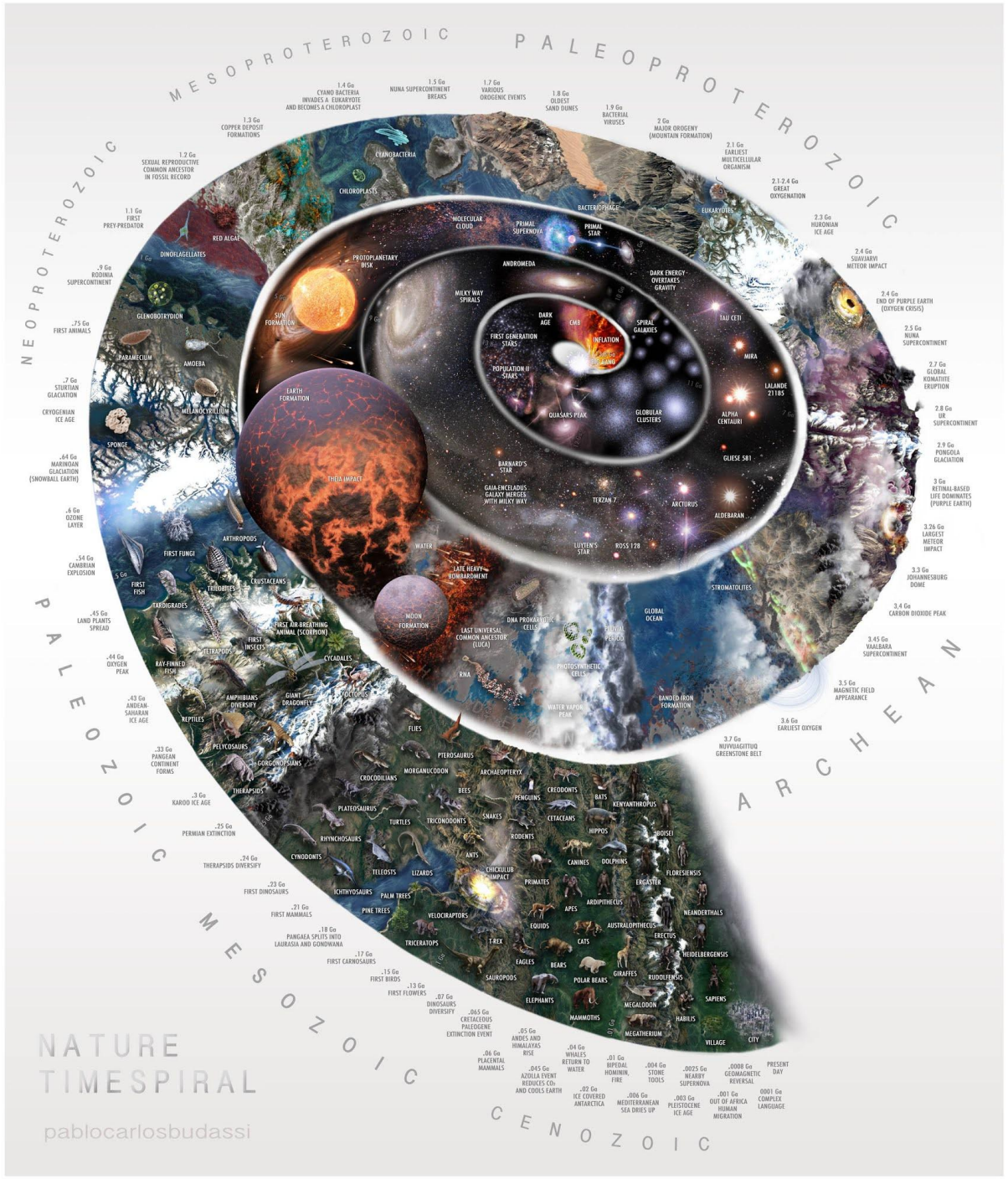
Article | Published: 22 December 2021

A species-level timeline of mammal evolution integrating phylogenomic data

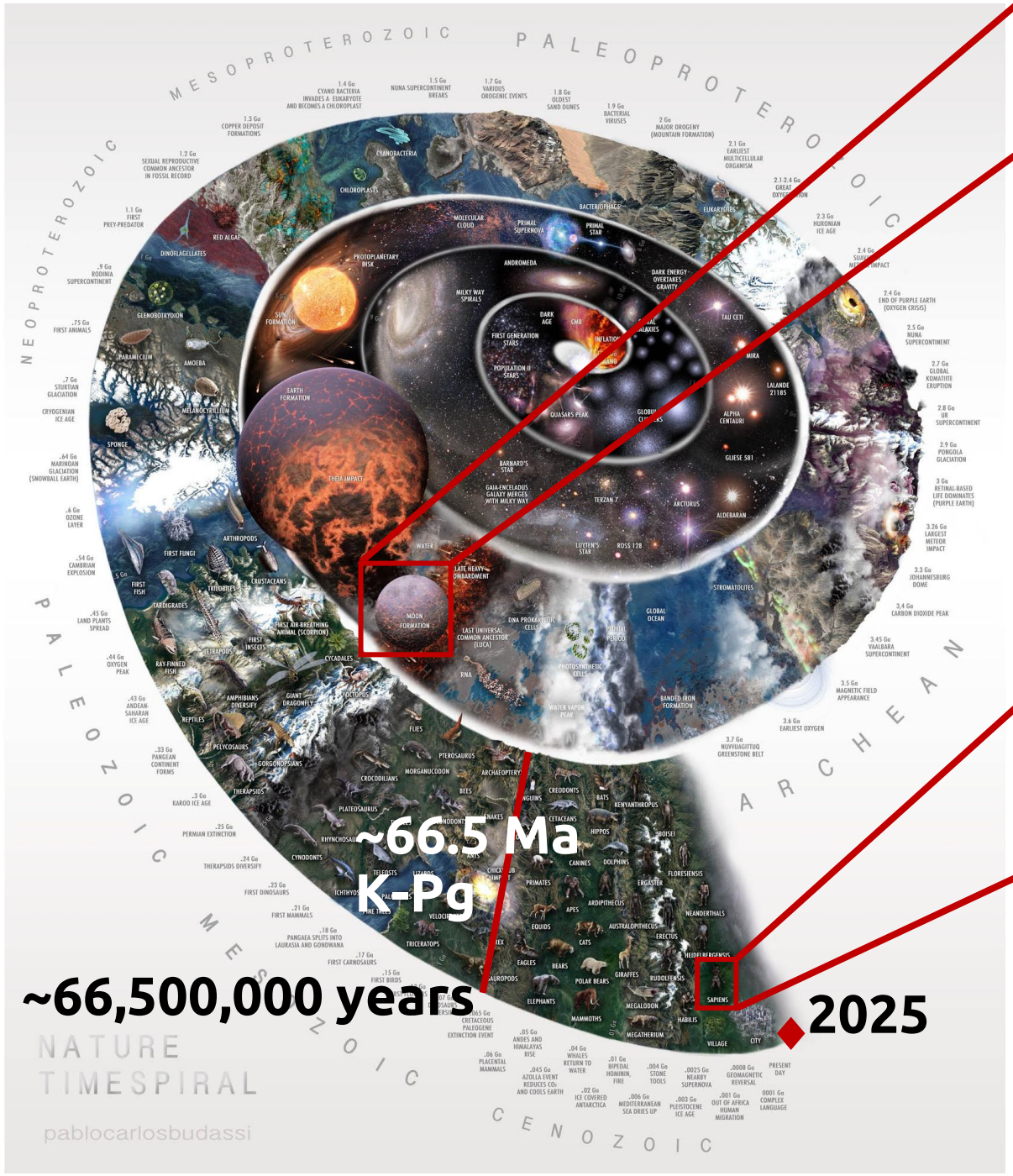
Sandra Álvarez-Carretero, Asif U. Tamuri, Matteo Battini, Fabrícia F. Nascimento, Emily Carlisle, Robert J. Asher, Ziheng Yang, Philip C. J. Donoghue ✉ & Mario dos Reis ✉

Nature **602**, 263–267 (2022)

Test macroevolutionary hypotheses



Evolutionary timelines can help us understand Earth's evolutionary history!



Moon's formation

~4.5 Ga

~4,500,000,000 years

LUCA(*)

Last Universal Common Ancestor

~4.2 Ga

~4,200,000,000 years

The nature of the last universal common ancestor and its impact on the early Earth system

Edmund R. R. Moody, Sandra Álvarez-Carretero, Tara A. Mahendrarajah, James W. Clark, Holly C. Betts, Nina Dombrowski, Lénárd L. Szánthó, Richard A. Boyle, Stuart Daines, Xi Chen, Nick Lane, Ziheng Yang, Graham A. Shields, Gergely J. Szöllősi, Anja Spang, Davide Pisani, Tom A. Williams, Timothy M. Lenton & Philip C. J. Donoghue

Nature Ecology & Evolution **8**, 1654–1666 (2024)

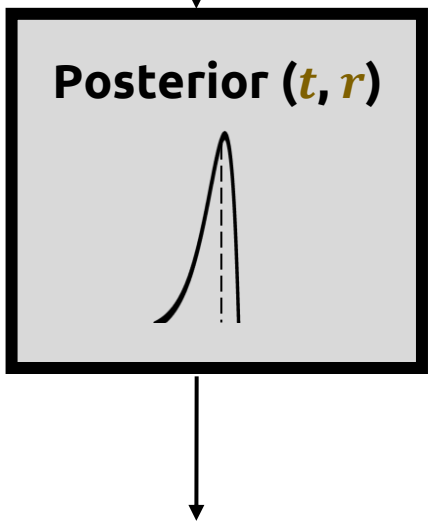
Homo sapiens

~300,000 years

Evolutionary timeline of LUCA

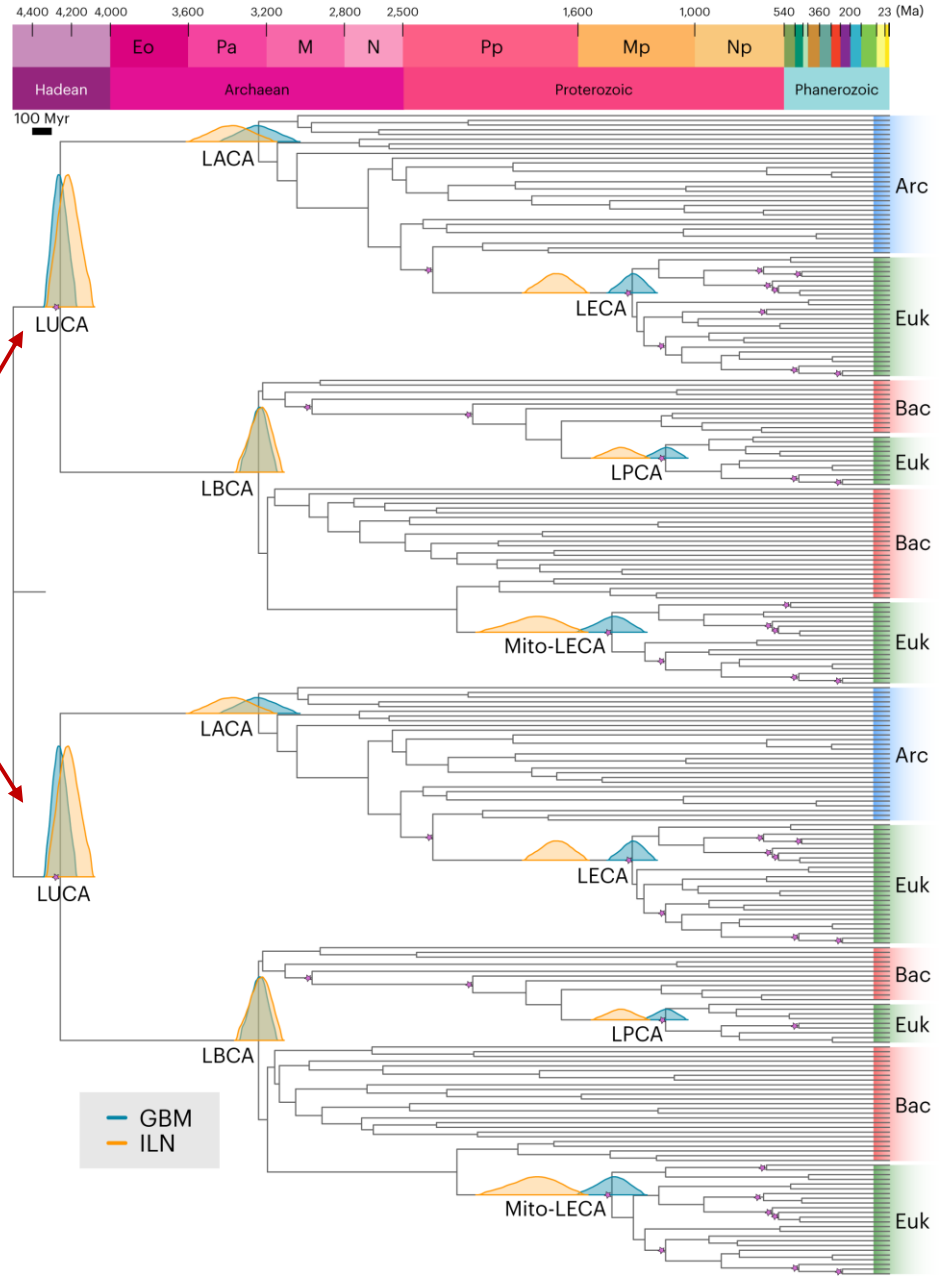
$$posterior \propto priors \times likelihood$$

$$f(t, r|D) \propto f(t) \times f(r|t) \times f(D|t, r)$$



Mean time estimate for LUCA

~4.2 Ga



Our results look like that: time probability distributions!
We take the mean of these distributions across all nodes in the phylogeny as well as the confidence intervals, which help us integrate the uncertainty of our time estimates!

GBM = Geometric Brownian Motion (Autocorrelated rates model)
ILN = Independent lognormal (Independent rates model)

Why are evolutionary timelines useful?

- The **molecular clock** has been key to understanding the relationship between evolutionary rates and divergence times.
- The **clock** only holds for **closely-related species**, and thus **relaxing** the clock is required for most current analyses with large genomic datasets.
- **Bayesian approaches** are the main chosen methods for **clock-dating analyses** given how easy it is to **integrate the uncertainty on model parameters** (e.g., rates and times) through the usage of **priors**.
- Studying **Earth's biodiversity** and **testing contentious macroevolutionary questions** within the fields of evolution, ecology, and even conservation are the main applications of **evolutionary timelines**.

Which software can we use to infer evolutionary timelines?

Software for timetree inference

- `MCMCtree` (part of PAML, Yang 2007).
- `McmcDate` (Schrempf et al. [unpublished]; see Harris et al. 2022 for first application).
- `PhyloBayes` (Lartillot and Philippe, 2004; but see also Lartillot 2020).
- `MrBayes` (Huelsenbeck and Ronquist, 2001).
- `BEAST` (Suchard et al., 2018) and `BEAST2` (Bouckaert et al., 2019).
- `RevBayes` (Höhna et al., 2016).

Software for timetree inference

- **MCMCtree** (part of PAML, Yang 2007).
- **McmcDate** (Schrempf et al. [unpublished]; see Harris et al. 2022 for first application).
- PhyloBayes (Lartillot and Philippe, 2004; but see also Lartillot 2020).
- MrBayes (Huelsenbeck and Ronquist, 2001).
- BEAST (Suchard et al., 2018) and BEAST2 (Bouckaert et al., 2019).
- RevBayes (Höhna et al., 2016).

MCMCtree and McmcDate are the only two software (at the time of writing) that have implemented an approximation to the likelihood calculation that enables large phylogenomic datasets to be analysed using a reasonable number of computational resources for a reasonable amount of time (e.g., from days to a month or few months, depending on data size).

NOTE: we will learn how to run MCMCtree during the next practical session!

SO…

How does this approximation work in MCMCtree?

Bayesian statistics applied to molecular-clock dating analyses

$$posterior \propto prior \times likelihood$$

$$f(t, r|D) \propto f(t) f(r|t) f(D|t, r)$$

WHY DOES IT TAKE SO LONG TO ESTIMATE THE POSTERIOR WITH PHYLOGENOMIC DATA?

D = molecular data
t = vector of divergence times
r = vector of molecular rates
θ = vector of other unknown parameter/s

Bayesian statistics applied to molecular-clock dating analyses

$$posterior \propto prior \times likelihood$$

$$f(t, r|D) \propto f(t) f(r|t) f(D|t, r)$$

BECAUSE THE TIME TO CALCULATE THE LIKELIHOOD IS PROPORTIONAL TO THE NUMBER OF SITE PATTERNS IN THE ALIGNMENT!

D = molecular data
$\boldsymbol{t}$ = vector of divergence times
$\boldsymbol{r}$ = vector of molecular rates
$\boldsymbol{\theta}$ = vector of other unknown parameter/s

Bayesian statistics applied to molecular-clock dating analyses

$$posterior \propto prior \times likelihood$$

$$f(t, r|D) \propto f(t) f(r|t) f(D|t, r)$$

HOW CAN WE SPEED THINGS UP WITH PHYLOGENOMIC DATA?
APPROXIMATE THE LIKELIHOOD CALCULATION

D = molecular data
$\boldsymbol{t}$ = vector of divergence times
$\boldsymbol{r}$ = vector of molecular rates
$\boldsymbol{\theta}$ = vector of other unknown parameter/s

Approximating the likelihood calculation

Approximate Likelihood Calculation on a Phylogeny for Bayesian Estimation of Divergence Times

Mario dos Reis, Ziheng Yang ✉ Author Notes

Molecular Biology and Evolution, Volume 28, Issue 7, July 2011, Pages 2161–2172, https://doi.org/10.1093/molbev/msr045

Published: 10 February 2011

Use Taylor expansion of the log-likelihood:

a) **Vector of branch lengths** (substitutions/site): $\mathbf{b} = \{b_i = t_i r_i\}$

Molecular rate on branch *i*: r_i

Time duration of *i*-th branch: t_i

b) **Log-likelihood as a function of branch lengths**: $l(\mathbf{b}) = \log f(D|\,\mathbf{t}, \mathbf{r})$

c) **Taylor expansion around MLEs of branch lengths:**

$$f(D|\mathbf{t}, \boldsymbol{r}) \rightarrow l(\mathbf{b}) \approx l(\hat{\mathbf{b}}) + \mathbf{g}^T(\mathbf{b} - \hat{\mathbf{b}}) + \frac{1}{2}\Delta\mathbf{b}^T\mathbf{H}(\mathbf{b} - \hat{\mathbf{b}})$$

MLEs branch lengths: $\hat{\mathbf{b}}$

Vector of first derivatives, **gradient**: $\mathbf{g} = \{g_i\}$

Matrix of second derivatives, **Hessian:** $\mathbf{H} = \{H_{ij}\}$

Approximating the likelihood calculation

Things to consider when approximating the likelihood calculation:

- Remove taxa and/or partitions with very long ("infinite") branches.
  Infinite values are outside the parameter space
- Re-estimate branch lengths, Hessian, and gradient if testing another tree topology.
- Co-estimation of tree topology and divergence times is not possible.

Approximate Likelihood Calculation on a Phylogeny for Bayesian Estimation of Divergence Times

Mario dos Reis, Ziheng Yang ✉ Author Notes

Molecular Biology and Evolution, Volume 28, Issue 7, July 2011, Pages 2161–2172, https://doi.org/10.1093/molbev/msr045

Published: 10 February 2011

Approximating the likelihood calculation

93
((((((((((((((tax_1: 0.349172, ((tax_2: 0.090297, tax_3: 0.100873): 0.009834, (tax_4: 0.173676, (tax_5: 0.087587 […]

$\hat{\mathbf{b}}$
0.006541 0.014905 0.017282 0.007034 0.006182 0.023338 0.012032 0.000748 0.035520 0.019296 0.010760 0.021452 […]

$\mathbf{g}$
0 0.030360 0.009750 -0.010872 0.011271 -0.013595 -0.008532 0.010353 0.007271 0.000000 -0.005711 -0.013245 -0.018805 […]

Hessian

$\mathbf{H}$
-1.335e+05 -5756 -4287 -3683 -3929 -3297 -3530 -705.1 -898.9 -1087 -88.77
-5756 -4.899e+04 -2.452e+04 -1.373e+04 -1.249e+04 -8949 -9603 -5567 -4159 -3205 -3696
-4287 -2.452e+04 -6.025e+04 -1.976e+04 -1.453e+04 -1.27e+04 -8261 -7463 -6004 -3632 -2125
-3683 -1.373e+04 -1.976e+04 -1.017e+05 -3.662e+04 -2.088e+04 -9936 -7531 -6141 -1.196e+04 -3645
[…]

CALCULATED BY BASEML (nuc) OR CODEML (prot)!

MLEs branch lengths: $\hat{\mathbf{b}}$

Vector of first derivatives, **gradient**: $\mathbf{g} = \{g_i\}$

Matrix of second derivatives, **Hessian**: $\mathbf{H} = \{H_{ij}\}$

Approximating the likelihood calculation

MCMCtree WILL USE THAT TO APPROXIMATE THE LIKELIHOOD CALCULATION!

$$f(D|\boldsymbol{t},\boldsymbol{r}) \rightarrow l(\mathbf{b}) \approx l(\hat{\mathbf{b}}) + \mathbf{g}^T \Delta\mathbf{b} + \frac{1}{2}\Delta\mathbf{b}^T \mathbf{H} \Delta\mathbf{b}$$

CALCULATED BY BASEML (nuc) OR CODEML (prot)!

MLEs branch lengths: $\hat{\mathbf{b}}$

Vector of first derivatives, **gradient**: $\mathbf{g} = \{g_i\}$

Matrix of second derivatives, **Hessian:** $\mathbf{H} = \{H_{ij}\}$

Time for questions

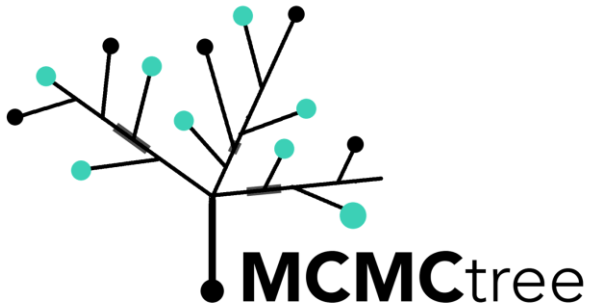
Let’s get ready for the practical session !

PAML GitHub:
https://github.com/abacus-gene/paml

PAML Wiki
PAML docs

BAYESIAN TIMETREE INFERENCE

with MCMCtree



Groups

PAML Discussion Group:
https://groups.google.com/g/pamlsoftware

Step 0: data formatting

To learn more and see other examples, please check:

PAML Wiki
PAML docs

To ease the analyses with `MCMCtree`, we need to format the raw data:

- Tree file:
  - **Calibrated tree (`MCMCtree`)**: **Newick** format without branch lengths or other types of labels except for the calibrations (e.g., soft bounds, skew-t, etc.).

```
4 1
(sp1,((sp2,sp3)'B(4.12,4.52)',sp4))'ST(5.83,0.059,0.112,109.124)’;
```

  - **Uncalibrated tree (`BASEML or CODEML`)**: **Newick** format without branch lengths or any type of label (i.e., just the tree topology).

```
4 1
(sp1,((sp2,sp3),sp4));
```

Step 0: data formatting

To learn more and see other examples, please check:

PAML Wiki
PAML docs

To ease the analyses with `MCMCtree`, we need to format the raw data:

- Types of calibrations (brief overview!):

Calibration	Notation
Lower/Minimum bound (L)	'>0.06' equals to **'L(0.06)'** * There are other notations
Upper/Maximum bound (U)	'>0.08' equals to **'U(0.08)'** *There are other notations
Lower+Upper/Min+Max bounds (B)	'>0.06<0.08' equals to **'B(0.06,0.08)'** *There are other notations
Gamma (G)	'G(alpha, beta)'
Sew normal (SN)	'SN(location, scale, shape)'
Skew t (ST)	'ST(location, scale, shape, df)'
S2N (Swek 2 normal)	'SN2(p1, loc1, scale1, shape1, locs2, scale2, shape2)'

To learn more and see other examples, please check:

PAML Wiki
PAML docs

Step 0: data formatting

To ease the analyses with `MCMCtree`, we need to format the raw data:

- Tree file:
  - **Calibrated tree (`MCMCtree`)**: **Newick** format without branch lengths or other types of labels except for the calibrations (e.g., soft bounds, skew-t, etc.).
  - **Uncalibrated tree (`BASEML or CODEML`)**: **Newick** format without branch lengths or any type of label (i.e., just the tree topology).

- Alignment file: **PHYLIP** format, one sequence per row/line.

```
4   609

sp1       TTTAGTGTGCTTATTAGGTTAGAATTATCGGCT […]
sp2       TTTAGTATGTTAATTAGATTAGAGTTGTCTGGC […]
sp3       TTTAGTTTATTGATAAGATTAGAGCTATCAGGA […]
sp4       TTTAGTGTGCTTATTAGGTTAGAATTATCGGCT […]
```

Step 1: calculating bl, gradient, and Hessian (BASEML template)

```
seed = -1                   * Seed number. If -1, use time stamp
seqfile = ALN               * Path to alignment file
treefile = TREE             * Path to tree file
mcmcfile = mcmc.txt         * Path to file where MCMC samples will be saved
outfile = out.txt           * Path to where output file will be saved

ndata = 1                   * Number of partitions in the alignment
seqtype = 0                 * 0: nucleotides; 1:codons; 2:Aas
usedata = 3                 * 0: no data (prior); 1:exact likelihood;
                            * 2:Approx lnL; 3:out.BV (in.BV)
clock = 1                   * 1: STR; 2: ILN; 3: GBM
model = 4                   * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
alpha = 0.5                 * alpha for gamma rates at sites
ncatG = 5                   * No. categories in discrete gamma
cleandata = 0               * remove sites with ambiguity data (1:yes, 0:no)?
BDparas = 1 1 0.1           * birth, death, sampling
rgene_gamma = ALPHA BETA    * gammaDir prior for rate for genes
sigma2_gamma = ALPHA BETA   * gammaDir prior for sigma^2 (for clock=2 or 3)
print = 1                   * 0: no mcmc sample; 1: everything except
                            * branch rates 2: everything
burnin = 100000             * Samples to discard as part of burn-in phase
sampfreq = 1000             * Sampling frequency
nsample = 20000             * Total number of samples to collect during the MCMC
```

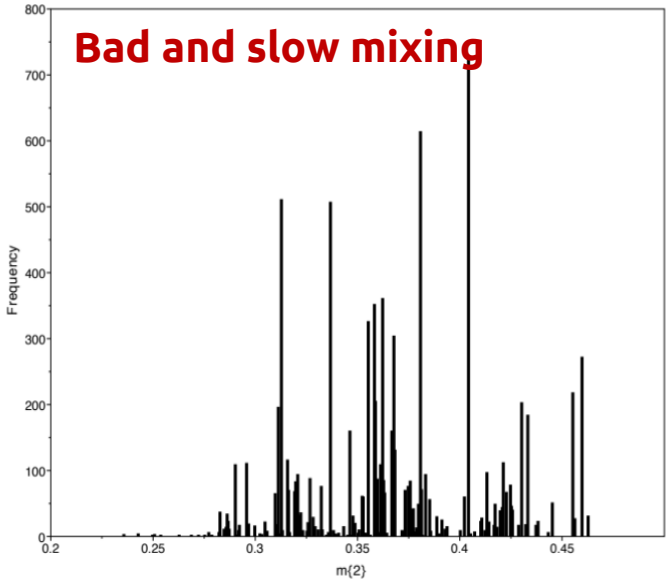
If running CODEML:

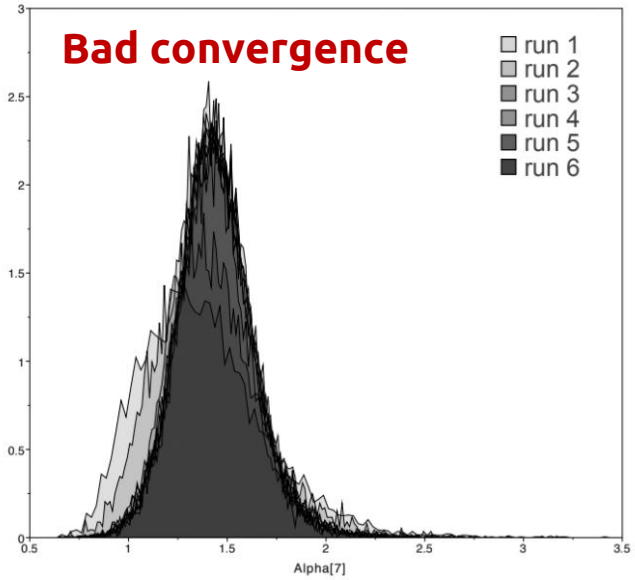
- **Add option `aaRatefile` with the path to the file with the rate matrix.**
- **Variable `model`:**
  - **`0:poisson`**
  - **`1:proportional`**
  - **`2:Empirical`**
  - **`3:Empirical+F`**
  - **`6:FromCodon`**
  - **`8:REVaa_0`**
  - **`9:REVaa(nr=189)`**

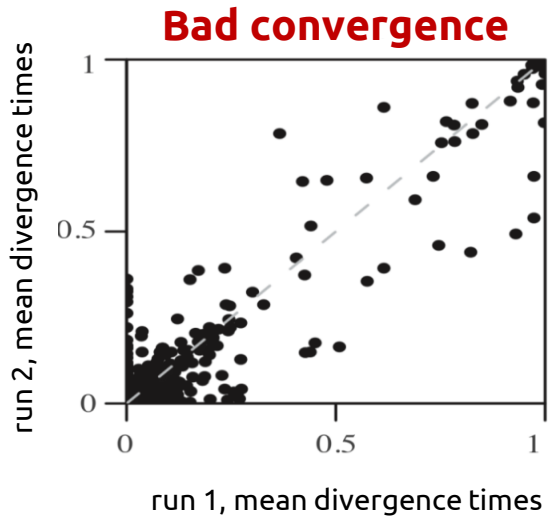
Step 2: running MCMCtree without data

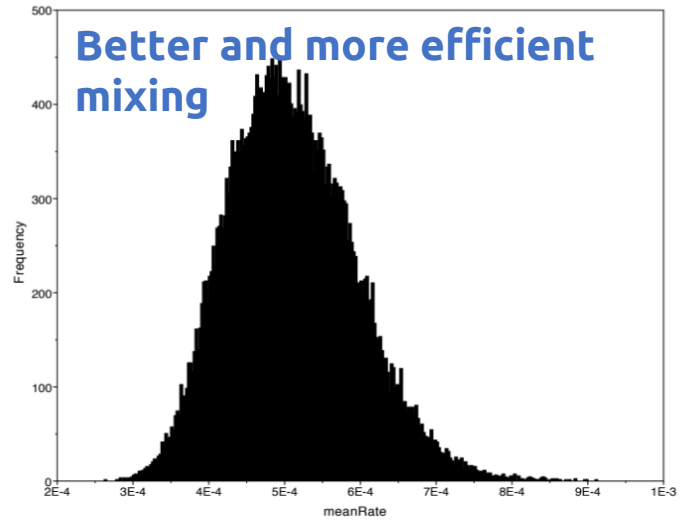
```
seed = -1                   * Seed number. If -1, use time stamp
seqfile = ALN               * Path to alignment file
treefile = TREE             * Path to tree file
mcmcfile = mcmc.txt         * Path to file where MCMC samples will be saved
outfile = out.txt           * Path to where output file will be saved

ndata = 1                   * Number of partitions in the alignment
seqtype = 0                 * 0: nucleotides; 1:codons; 2:Aas
usedata = 0                 * 0: no data (prior); 1:exact likelihood;
                            * 2:Approx lnL; 3:out.BV (in.BV)
clock = 1                   * 1: STR; 2: ILN; 3: GBM
model = 0                   * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
alpha = 0.5                 * alpha for gamma rates at sites
ncatG = 5                   * No. categories in discrete gamma
cleandata = 0               * remove sites with ambiguity data (1:yes, 0:no)?
BDparas = 1 1 0.1           * birth, death, sampling
rgene_gamma = ALPHA BETA    * gammaDir prior for rate for genes
sigma2_gamma = ALPHA BETA   * gammaDir prior for sigma^2 (for clock=2 or 3)
print = 1                   * 0: no mcmc sample; 1: everything except
                            * branch rates 2: everything
burnin = 100000             * Samples to discard as part of burn-in phase
sampfreq = 1000             * Sampling frequency
nsample = 20000             * Total number of samples to collect during the MCMC
```

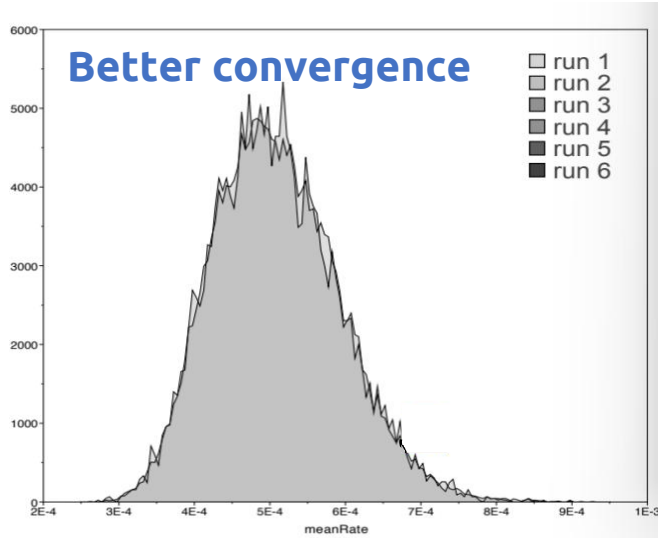
Step 3: assessing chain convergence and ESS

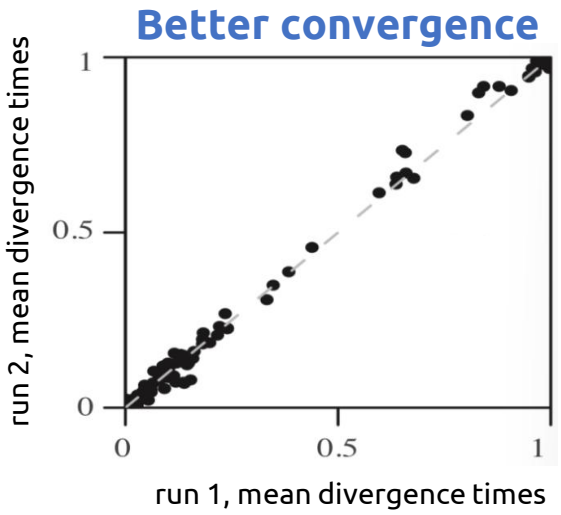




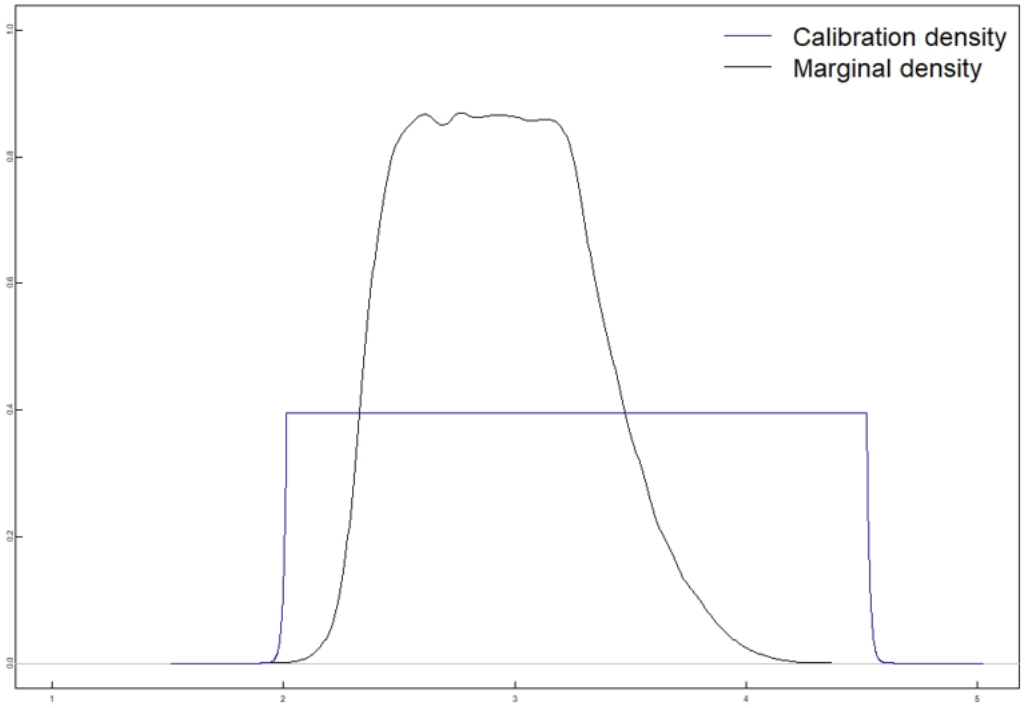


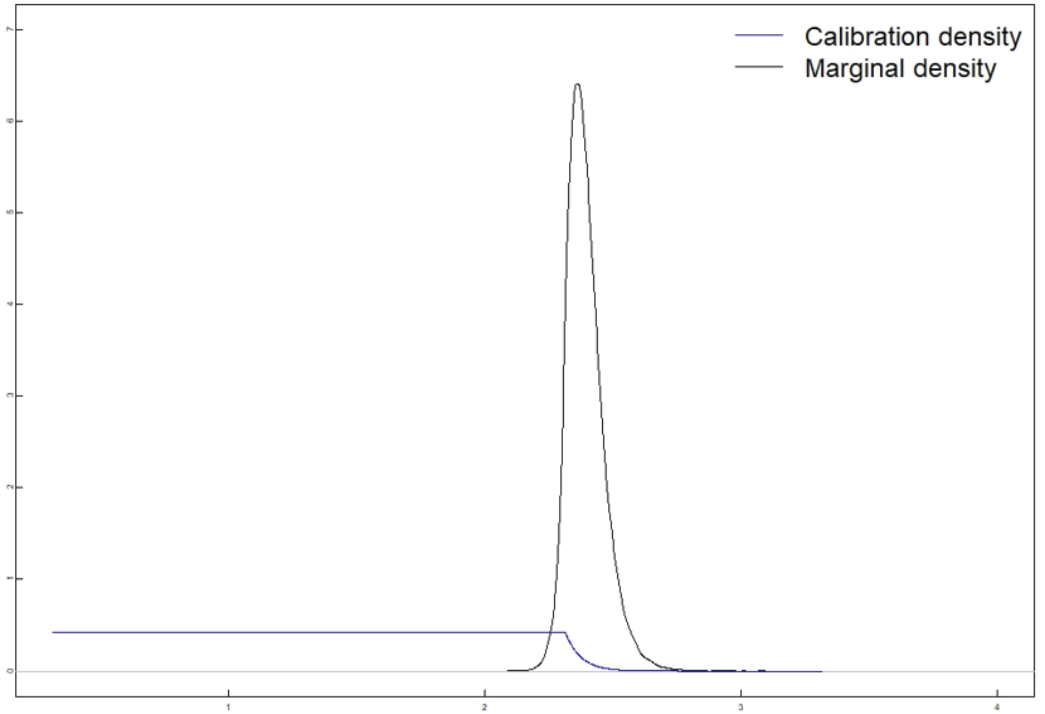






Step 4: comparing calibration densities VS marginal densities





Step 5: timetree inference with MCMCtree (approx. lnL)

```
seed = 1                  * Seed number. If -1, use time stamp
seqfile = ALN             * Path to alignment file
treefile = TREE           * Path to tree file
mcmcfile = mcmc.txt       * Path to file where MCMC samples will b
outfile = out.txt         * Path to where output file will be save

ndata = 1                 * Number of partitions in the alignment
seqtype = 0               * 0: nucleotides; 1:codons; 2:Aas
usedata = 2 ./in.BV       * 0: no data (prior); 1:exact likelihood
                          * 2:Approx lnL; 3:out.BV (in.BV)
clock = 3                 * 1: STR; 2: ILN; 3: GBM
model = 4                 * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
alpha = 0.5               * alpha for gamma rates at sites
ncatG = 5                 * No. categories in discrete gamma
cleandata = 0             * remove sites with ambiguity data (1:ye
BDparas = 1 1 0.1         * birth, death, sampling
rgene_gamma = ALPHA BETA  * gammaDir prior for rate for genes
sigma2_gamma = 1 10       * gammaDir prior for sigma^2 (for clock=
print = 1                 * 0: no mcmc sample; 1: everything excep
                          * branch rates 2: everything
burnin = 100000           * Samples to discard as part of burn-in
sampfreq = 1000           * Sampling frequency
nsample = 20000           * Total number of samples to collect dur
```

If running the exact likelihood (i.e., Felsenstein's approach):

- **You will not run BASEML or CODEML (ignore step 1).**
- **After step 0 and steps 2-4, go to step 5, but set `usedata = 1`.**
- **You will need to add two variables in the control file after BDparas:**
  - **`kappa_gamma` (prior on transition/transversion ratio, κ). Requires ALPHA and BETA as `rgene_gamma` & `sigma2_gamma`.**
  - **`alpha_gamma` (prior on α, gamma shape parameter for variable rates among sites). Requires ALPHA and BETA as `rgene_gamma` & `sigma2_gamma`.**
- **You will run only `MCMCtree`, feasible with short alignments (will take longer than the approx. method).**

Step 5: timetree inference with MCMCtree (approx. lnL)

```
seed = 1                    * Seed number. If -1, use time stamp
seqfile = ALN               * Path to alignment file
treefile = TREE             * Path to tree file
mcmcfile = mcmc.txt         * Path to file where MCMC samples will b
outfile = out.txt           * Path to where output file will be save

ndata = 1                   * Number of partitions in the alignment
seqtype = 0                 * 0: nucleotides; 1:codons; 2:Aas
usedata = 2 ./in.BV         * 0: no data (prior); 1:exact likelihood;
                            * 2:Approx lnL; 3:out.BV (in.BV)
clock = 3                   * 1: STR; 2: ILN; 3: GBM
model = 4                   * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
alpha = 0.5                 * alpha for gamma rates at sites
ncatG = 5                   * No. categories in discrete gamma
cleandata = 0               * remove sites with ambiguity data (1:yes, 0:no)?
BDparas = 1 1 0.1           * birth, death, sampling
rgene_gamma = ALPHA BETA    * gammaDir prior for rate for genes
sigma2_gamma = 1 10         * gammaDir prior for sigma^2 (for clock=2 or 3)
print = 1                   * 0: no mcmc sample; 1: everything except
                            * branch rates 2: everything
burnin = 100000             * Samples to discard as part of burn-in phase
sampfreq = 1000             * Sampling frequency
nsample = 20000             * Total number of samples to collect during the MCMC
```

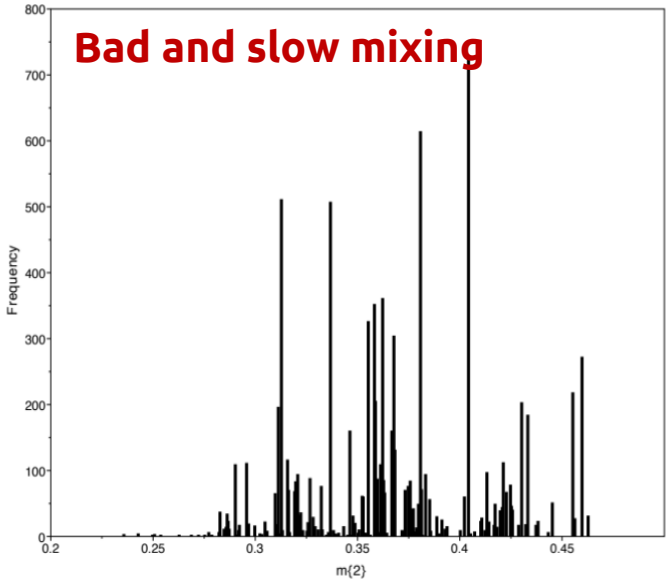
```
93
((((((((((((((((tax_1: 0.349172, ((tax_2: 0.090297, tax_3: 0.100873): 0.009834, (tax_4: 0.173676,

0.006541  0.014905  0.017282  0.007034  0.006182  0.023338  0.012032  0.000748  0.035520  0.019296[...]
 0  0.030360  0.009750 -0.010872  0.011271 -0.013595 -0.008532  0.010353  0.007271  0.000000 -0.005711 - [...]


Hessian
-1.335e+05    -5756    -4287    -3683    -3929    -3297    -3530    -705.1    -898.9    -1087    -88.77
-5756 -4.899e+04 -2.452e+04 -1.373e+04 -1.249e+04    -8949    -9603    -5567    -4159    -3205    -3696
-4287 -2.452e+04 -6.025e+04 -1.976e+04 -1.453e+04 -1.27e+04    -8261    -7463    -6004    -3632    -2125
-3683 -1.373e+04 -1.976e+04 -1.017e+05 -3.662e+04 -2.088e+04    -9936    -7531    -6141 -1.196e+04    -3645
[...]
```

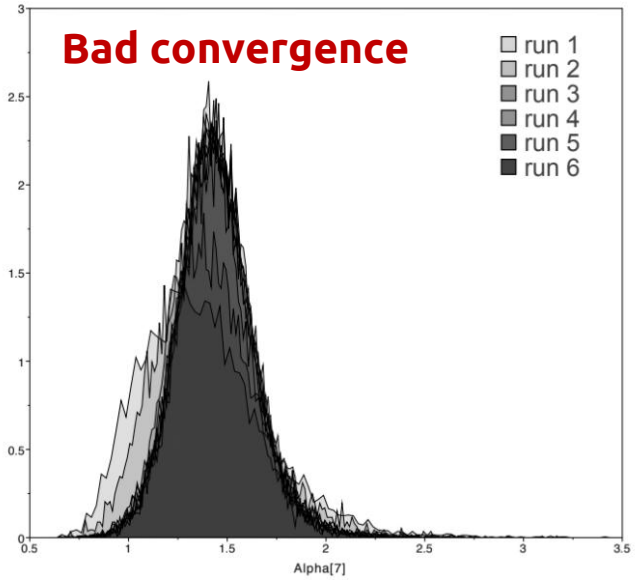
$\hat{\mathbf{b}}$ $\mathbf{g}$ $\mathbf{H}$

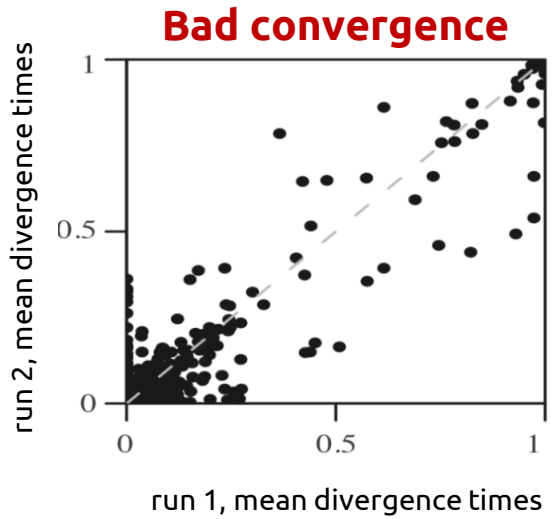
$$l(\mathbf{b}) \approx l(\hat{\mathbf{b}}) + \mathbf{g}^T \Delta\mathbf{b} + \frac{1}{2}\Delta\mathbf{b}^T \mathbf{H} \Delta\mathbf{b}$$

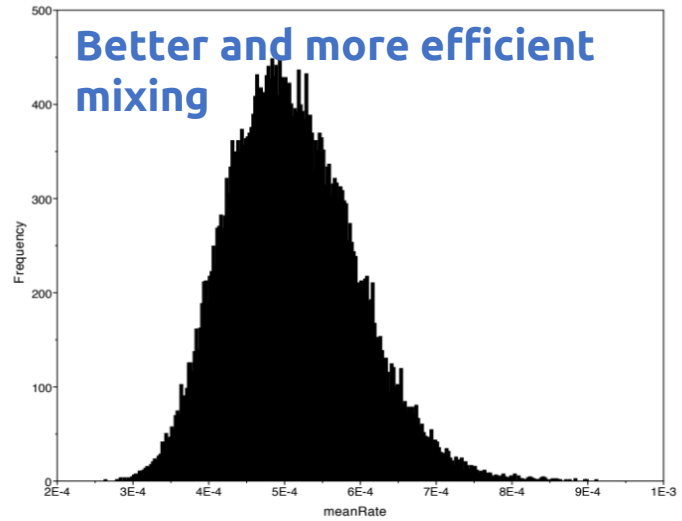
Options `model`, `alpha`, and `ncatG` are ignored if an `in.BV` file is used – BASEML/CODEML estimated the branch lengths, the gradient, and the Hessian under these settings already!

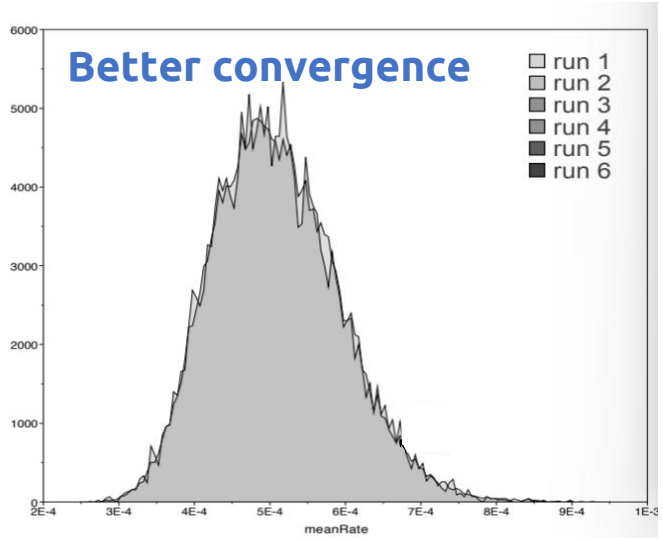
Step 6: assessing chain convergence and ESS

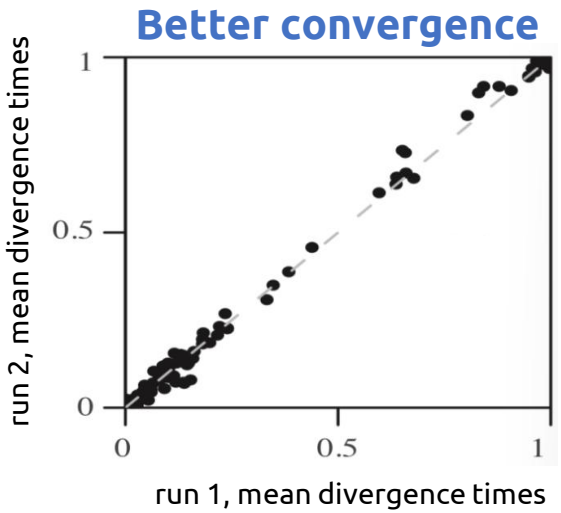












LET'S DO THIS!

https://github.com/abacus-gene/paml-tutorial/tree/main/mcmctree-approxlnL-aa

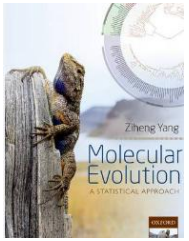
1. Clone and save the repository in your preferred location both in your PC and the server:
   `git clone https://github.com/abacus-gene/paml-tutorial/`

2. If you go inside the new cloned repository (`cd paml-tutorial`) and type `ls`, you will see various folders for different tutorials. Please access folder `mcmctree-approxlnL-aa` by typing `cd mmctree-approxlnL-aa` to check the tutorial for today's practical session.

3. If you had already cloned it, please go to directory `paml-tutorial` and type `git pull` to update the content – just in case there have been some changes in the code since the last time you cloned the repository!

4. Lastly, follow the README.md from your laptop (e.g., text or source code editor such as Visual Studio Code) or from the web browser and… Happy timetree inference!

Further reading



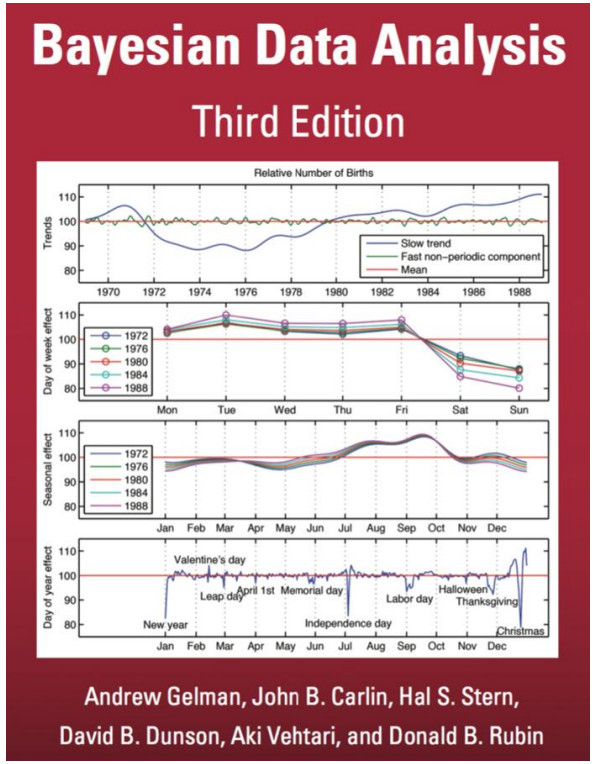
- "Molecular Evolution: a statistical approach", Yang (2014); see chapter 10.

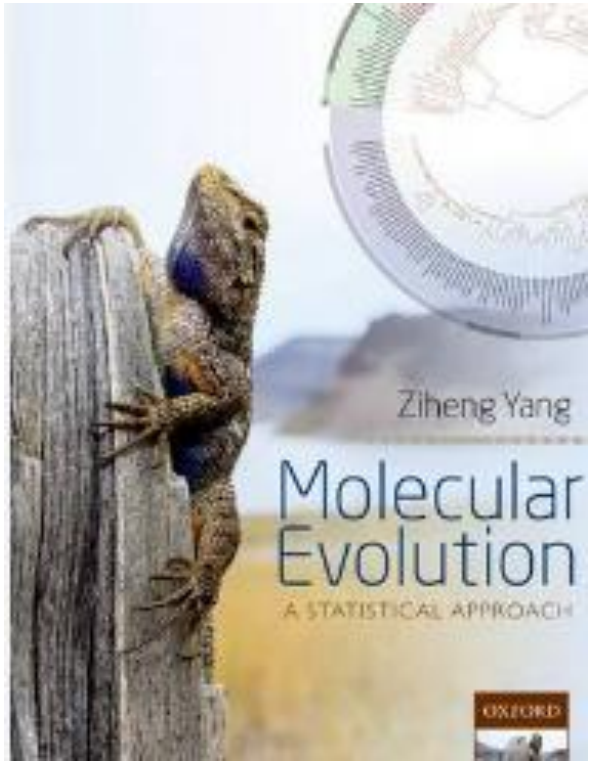
- Check Nascimento, dos Reis, and Yang (2017, *Nat Ecol Evol*, 1:1446-1454) for more details on Bayesian phylogenetic analyses and MCMC diagnostics.

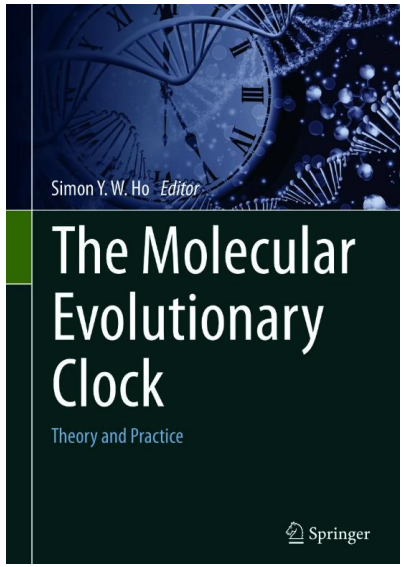
- For a general review on molecular clock-dating in the genomics era, please read dos Reis M, Donoghue PCJ and Yang Z. (2016) Nature Reviews Genetics, 17: 71–80.

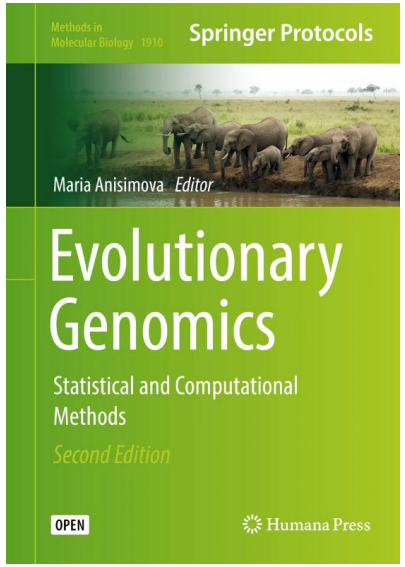
- For a review on Bayesian phylogenomic dating, please read Álvarez-Carretero S, and dos Reis M. (2021) In: Ho S (ed.) The Molecular Evolutionary Clock: Theory and Practice. Springer.

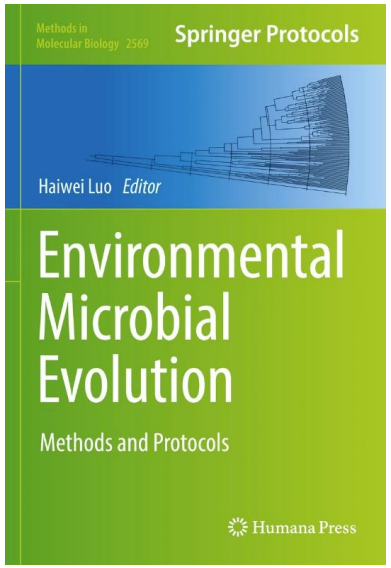
Books!

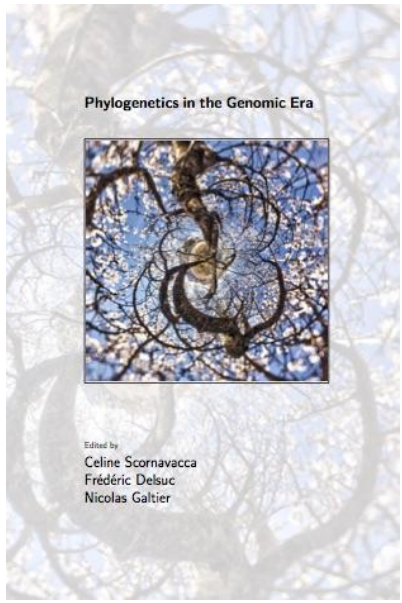

Bayesian Data Analysis
Third Edition
Relative Number of Births
Trends
Slow trend
Fast non-periodic component
Mean
Day of week effect
Seasonal effect
Day of year effect
Valentine's day
Leap day
April 1st
Memorial day
Labor day
Halloween
Thanksgiving
New year
Independence day
Christmas
Andrew Gelman, John B. Carlin, Hal S. Stern,
David B. Dunson, Aki Vehtari, and Donald B. Rubin

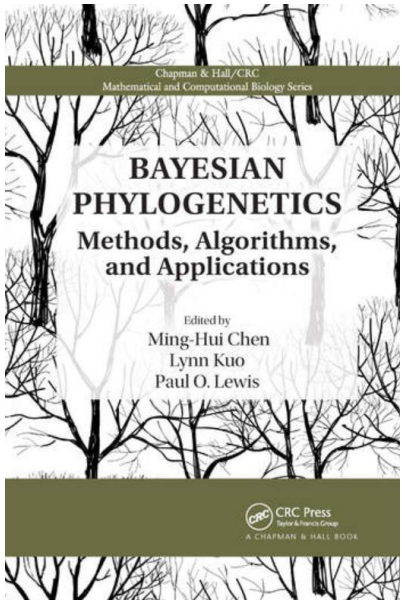

Ziheng Yang
Molecular Evolution
A STATISTICAL APPROACH
OXFORD

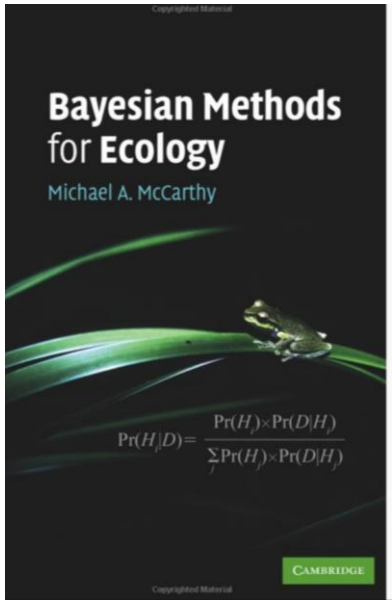

Simon Y. W. Ho Editor
The Molecular Evolutionary Clock
Theory and Practice
Springer


Methods in Molecular Biology 1910
Springer Protocols
Maria Anisimova Editor
Evolutionary Genomics
Statistical and Computational Methods
Second Edition
OPEN
Humana Press


Methods in Molecular Biology 2569
Springer Protocols
Haiwei Luo Editor
Environmental Microbial Evolution
Methods and Protocols
Humana Press

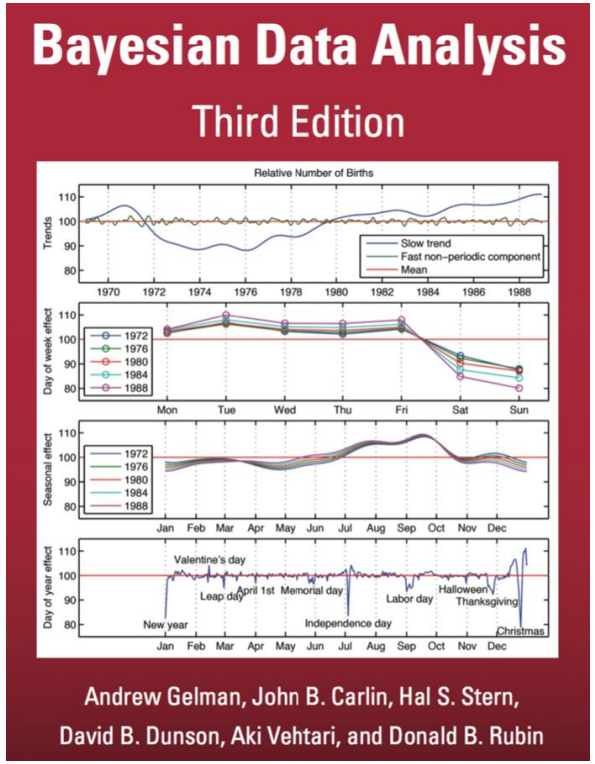

Phylogenetics in the Genomic Era
Edited by
Celine Scornavacca
Frédéric Delsuc
Nicolas Galtier

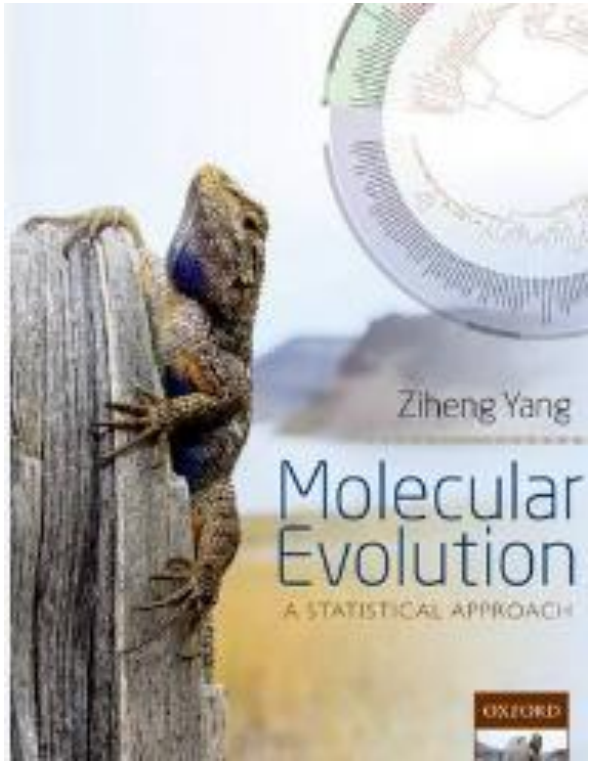

Chapman & Hall/CRC
Mathematical and Computational Biology Series
BAYESIAN PHYLOGENETICS
Methods, Algorithms, and Applications
Edited by
Ming-Hui Chen
Lynn Kuo
Paul O. Lewis
CRC Press
A CHAPMAN & HALL BOOK

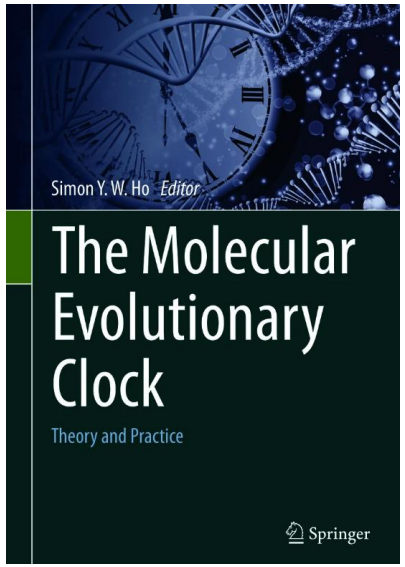

Bayesian Methods for Ecology
Michael A. McCarthy
CAMBRIDGE

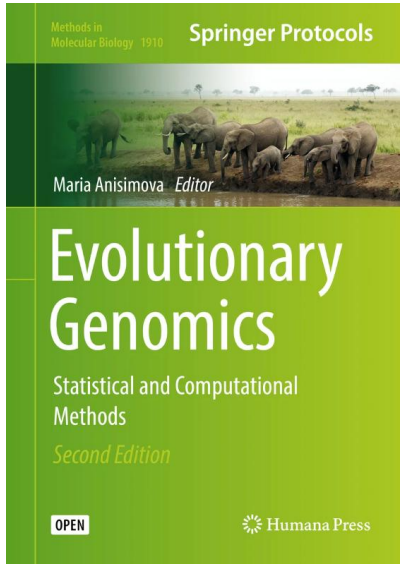
Books!

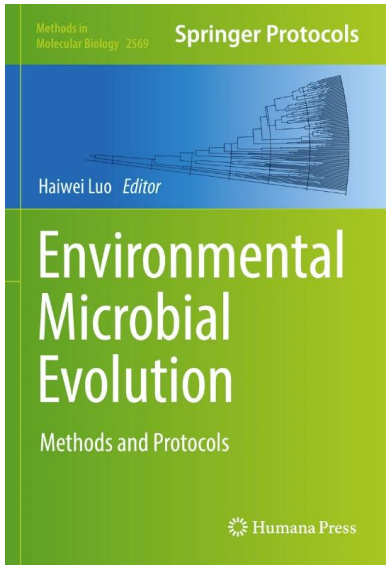
And I am sure that you might find many more!

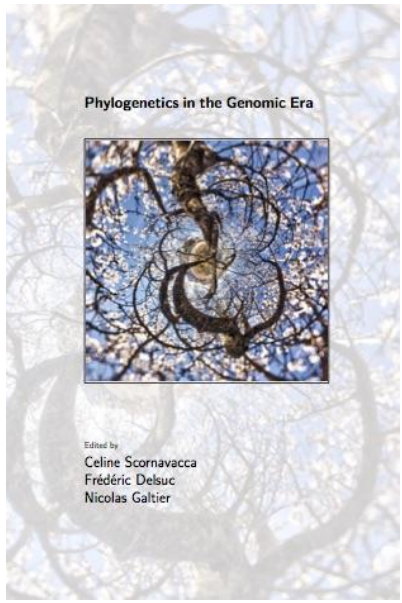


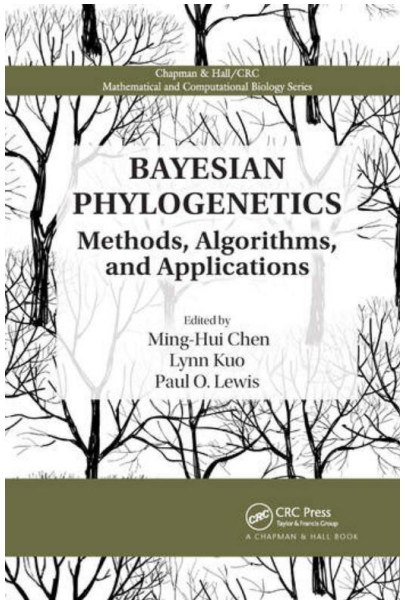


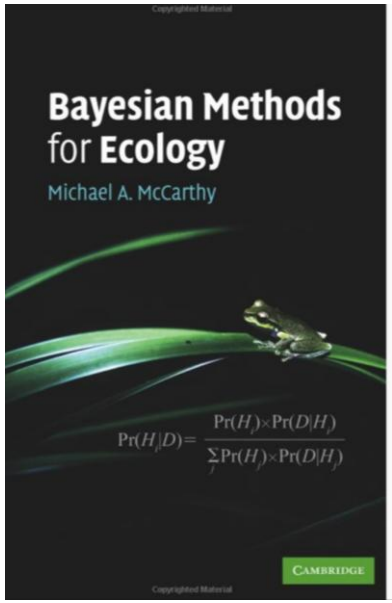












Resources consulted to generate these slides

CONTENT

I have assembled these slides by…

- … adapting material and resources taught in this module in previous years.
- … reusing material from previous seminars/workshops I have taught and/or created from scratch for this lecture.
- … consulting Prof Yang's book and the resources I was given while a participant at the CoME workshop in 2017 (Hinxton).

IMAGES

Images used are…

- … drawn/designed by me using Power Point or generated in R.
- … reused from previous material and/or extracted from cited papers and/or sites.
- … a combination of the above.