Clinical and Epidemiological Virology,
Rega Institute, Department of Microbiology
and Immunology
KU Leuven, Belgium.

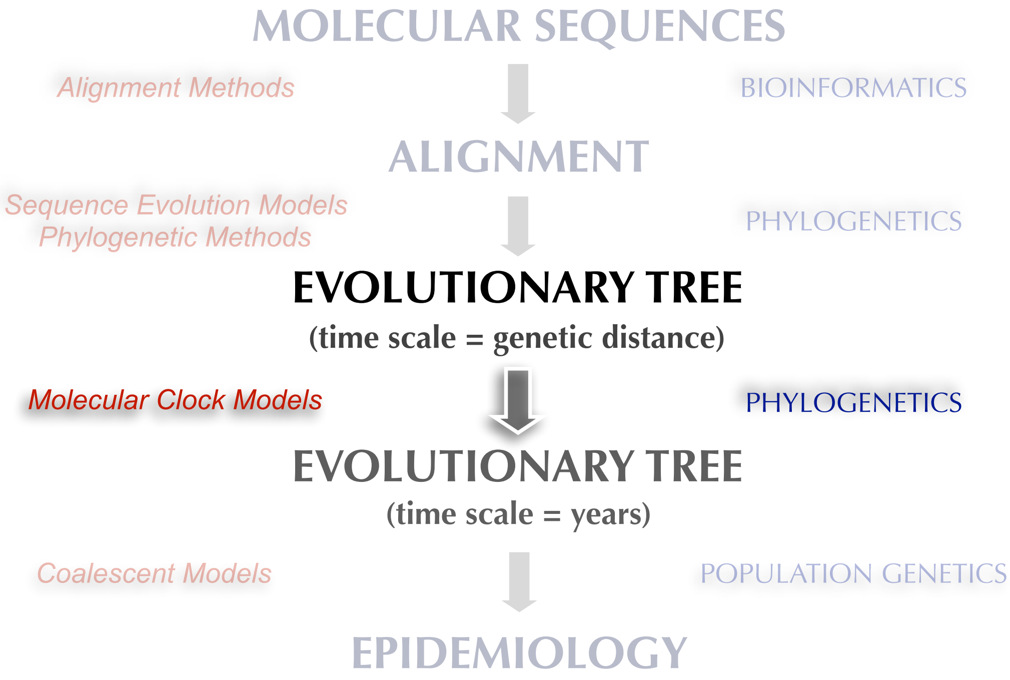
Estimating evolutionary rates and divergence times….

…and a bit of model testing

Philippe Lemey[1] and Marc Suchard[2]

1. Rega Institute, Department of Microbiology and Immunology, K.U. Leuven, Belgium.
2. Departments of Biomathematics and Human Genetics, David Geffen School of Medicine at UCLA. Department of Biostatistics, UCLA School of Public Health

SISMID, July 10-12, 2019



MOLECULAR SEQUENCES

Alignment Methods — BIOINFORMATICS

ALIGNMENT

Sequence Evolution Models
Phylogenetic Methods — PHYLOGENETICS

EVOLUTIONARY TREE
(time scale = genetic distance)

Molecular Clock Models — PHYLOGENETICS

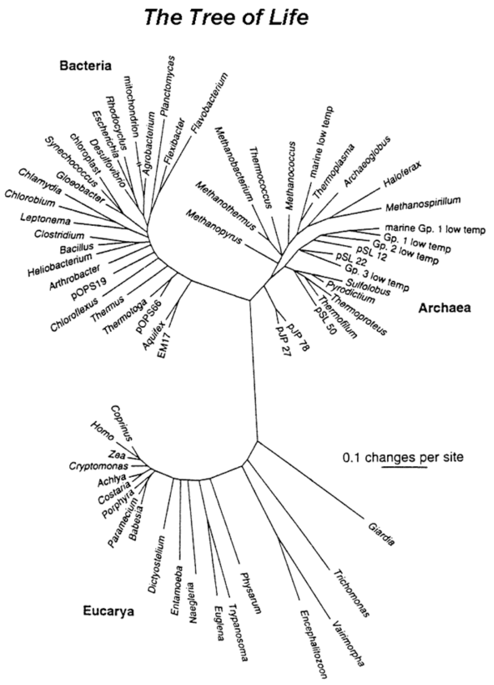
EVOLUTIONARY TREE
(time scale = years)

Coalescent Models — POPULATION GENETICS

EPIDEMIOLOGY

Molecular phylogenies

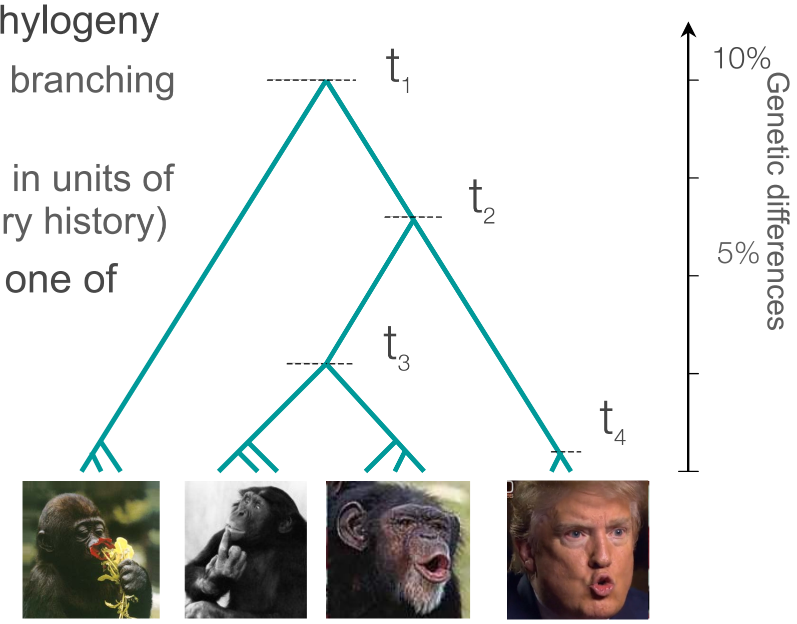
- most molecular phylogenies
  - are unrooted (or the rooting is due to prior information)
  - have branch lengths representing genetic change



- NR Pace, Science 1997

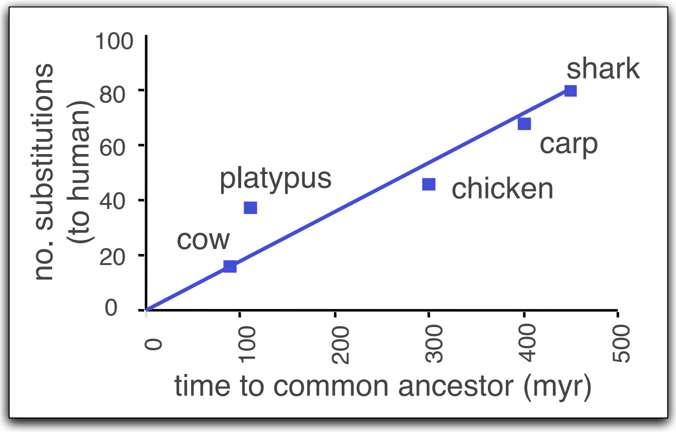
Molecular phylogenies

- the ideal molecular phylogeny
  - is rooted (implies a branching order)
  - has branch lengths in units of time (an evolutionary history)
- how do we construct one of these trees?



A constant evolutionary rate through time

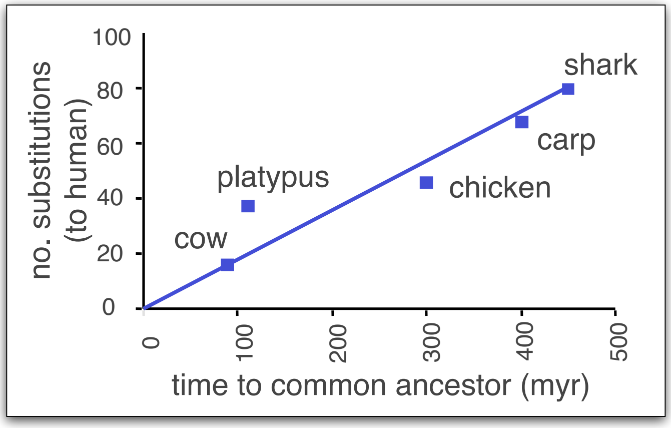
- to obtain a timed phylogeny, the evolutionary model must assume a relationship between the accumulation of genetic diversity and time



- Zuckerkandl and Pauling (1962): the rate of amino acid replacements in animal haemoglobins was roughly proportional to real time, as judged against the fossil record

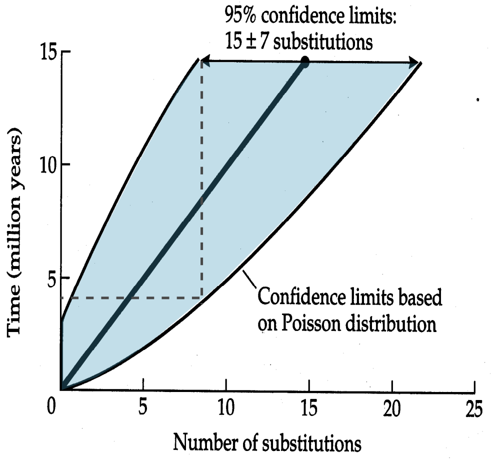
A constant evolutionary rate through time

- the *molecular clock* is particularly striking when compared to the obvious differences in rates of morphological evolution...



The molecular clock is not a metronome

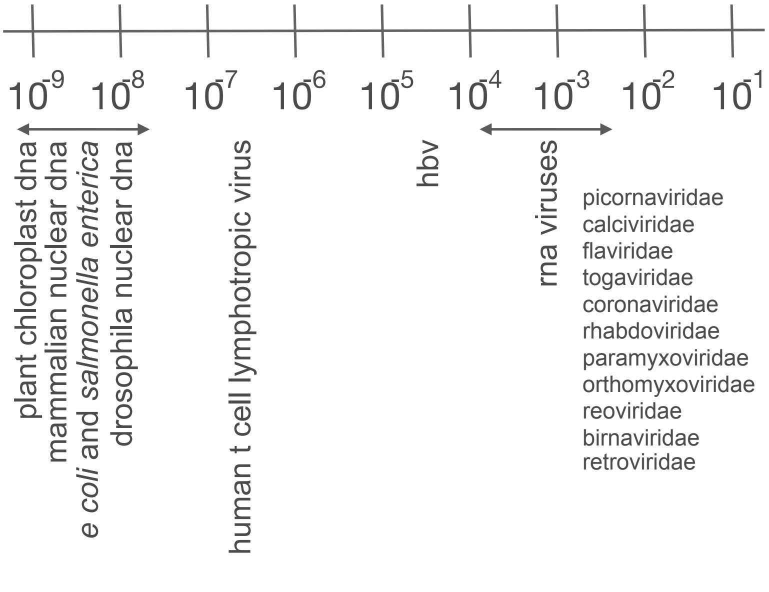
- if mutation every MY with Poisson variance
  - 95% of the lineages 15MY old have 8-22 substitutions
  - 8 substitutions also could be < 5 MY old



- Molecular Systematics, p532.

And there is no global molecular clock

nucleotide substitutions per site per year



plant chloroplast dna
mammalian nuclear dna
e coli and *salmonella enterica*
drosophila nuclear dna

human t cell lymphotropic virus

hbv

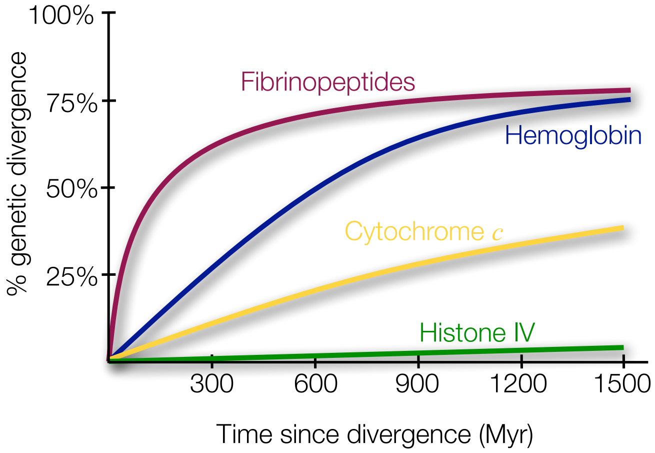
rna viruses

picornaviridae
calciviridae
flaviridae
togaviridae
coronaviridae
rhabdoviridae
paramyxoviridae
orthomyxoviridae
reoviridae
birnaviridae
retroviridae

And there is no global molecular clock

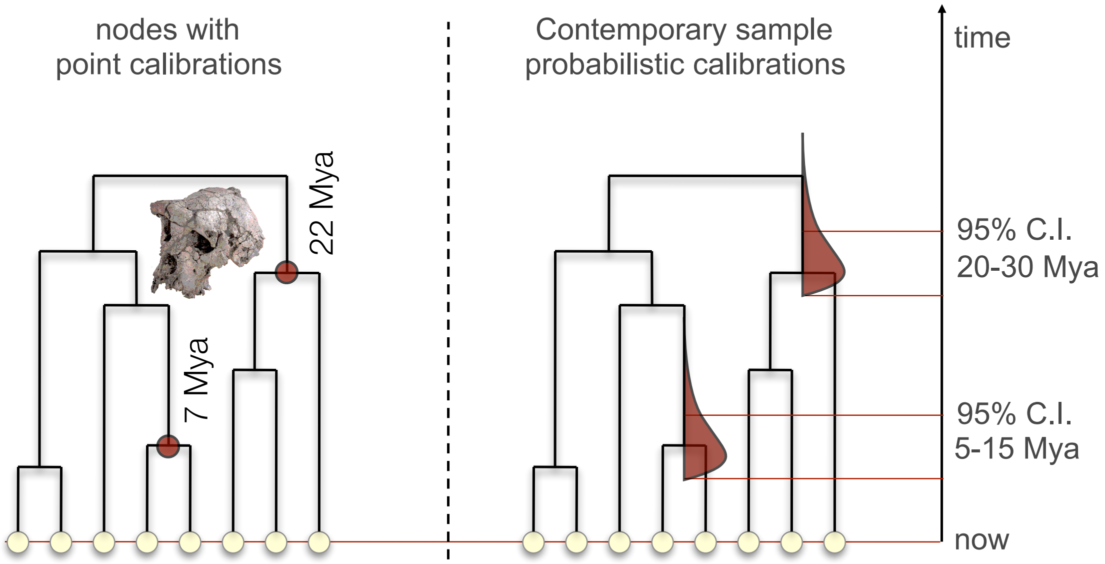
- different genes, different profiles
- variation in mutation rate?
- variation in selection

  genes coding for some molecules under very strong stabilizing selection



calibrating the molecular clock

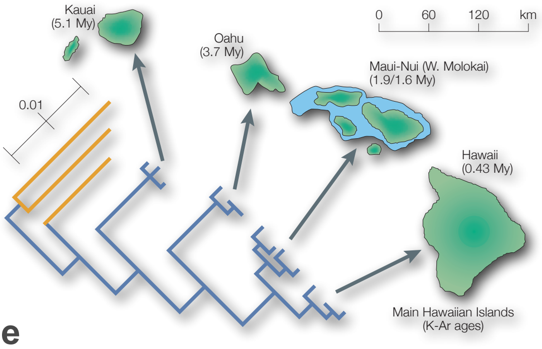
From substitution units to time units

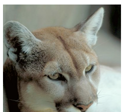


Node Calibrations

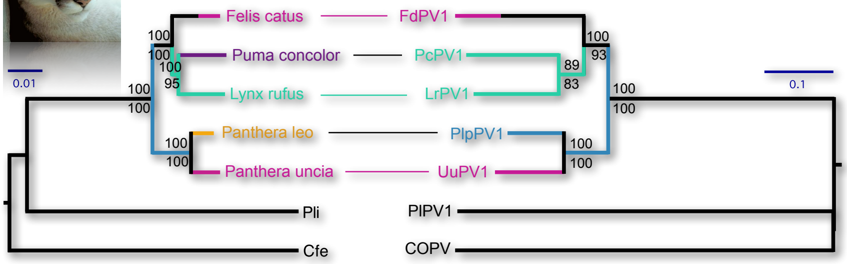
Fossils

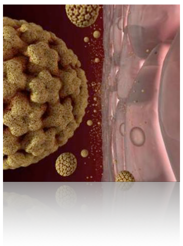
biogeography



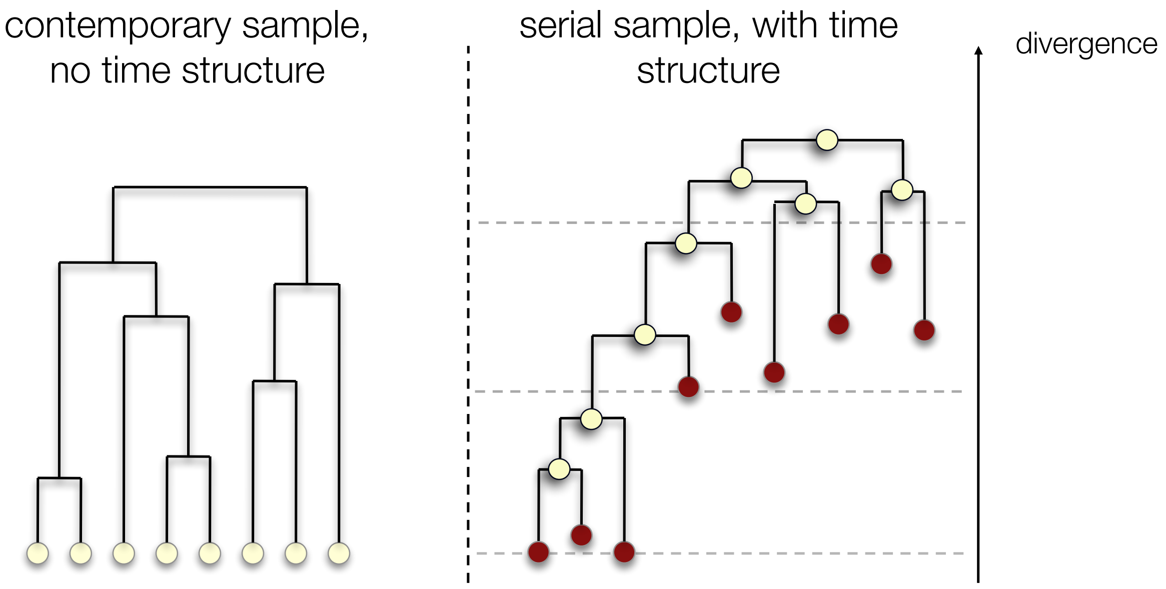


host-pathogen co-divergence

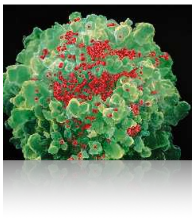




Calibration using sampling times



Tip calibration: two major applications

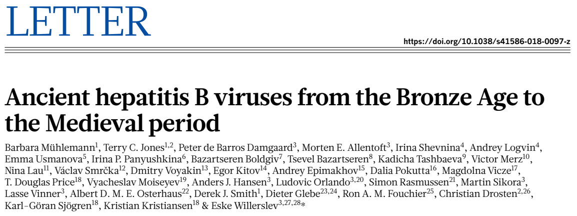


RNA viruses evolve quickly: 10^{-3} - 10^{-5} substitutions per site per year.

ancient DNA data sets of radiocarbon-dated specimens

- Substitutions accumulate between the times of sampling
- Serially sampled sequences or heterochronous sequences

Measurably evolving population



LETTER

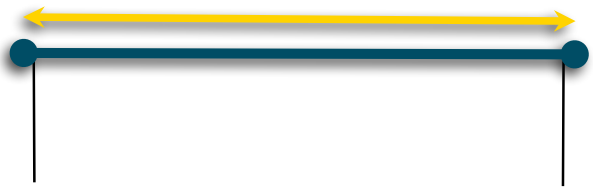
https://doi.org/10.1038/s41586-018-0097-z

Ancient hepatitis B viruses from the Bronze Age to the Medieval period

Barbara Mühlemann[1], Terry C. Jones[1,2], Peter de Barros Damgaard[3], Morten E. Allentoft[3], Irina Shevnina[4], Andrey Logvin[4], Emma Usmanova[5], Irina P. Panyushkina[6], Bazartseren Boldgiv[7], Tsevel Bazartseren[8], Kadicha Tashbaeva[9], Victor Merz[10], Nina Lau[11], Václav Smrčka[12], Dmitry Voyakin[13], Egor Kitov[14], Andrey Epimakhov[15], Dalia Pokutta[16], Magdolna Vicze[17], T. Douglas Price[18], Vyacheslav Moiseyev[19], Anders J. Hansen[3], Ludovic Orlando[3,20], Simon Rasmussen[21], Martin Sikora[3], Lasse Vinner[3], Albert D. M. E. Osterhaus[22], Derek J. Smith[1], Dieter Glebe[23,24], Ron A. M. Fouchier[25], Christian Drosten[2,26], Karl-Göran Sjögren[18], Kristian Kristiansen[18] & Eske Willerslev[3,27,28]*

incorporating sampling time: naive method

observed number of substitutions
or genetic divergence

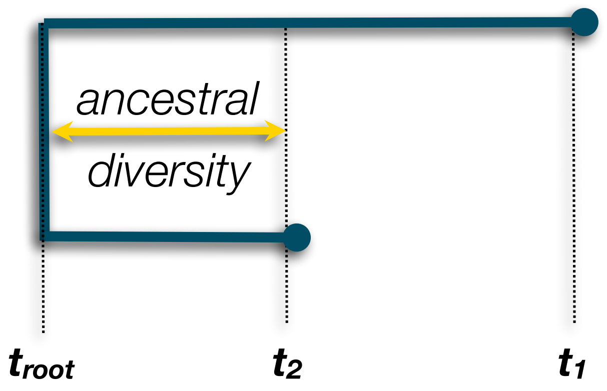


sampling time 1
t_1

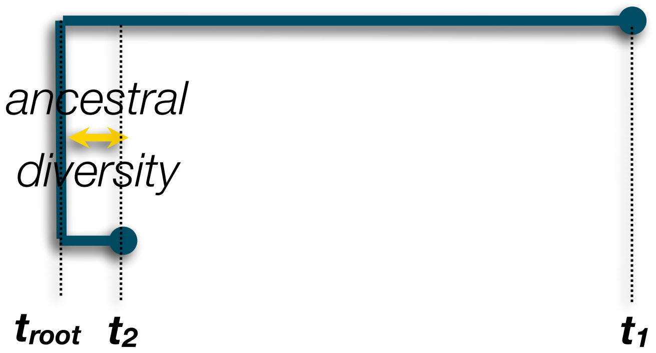
sampling time 2
t_2

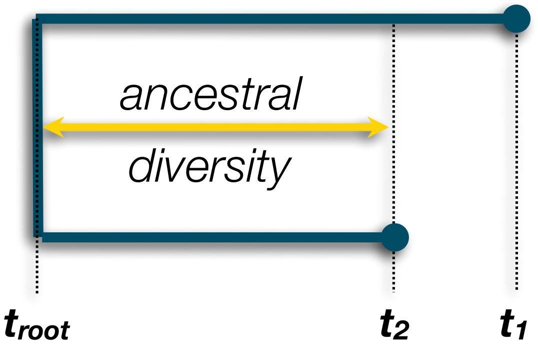
substitution rate, μ
$= d / |t_1 - t_2|$

incorporating sampling time: naive method

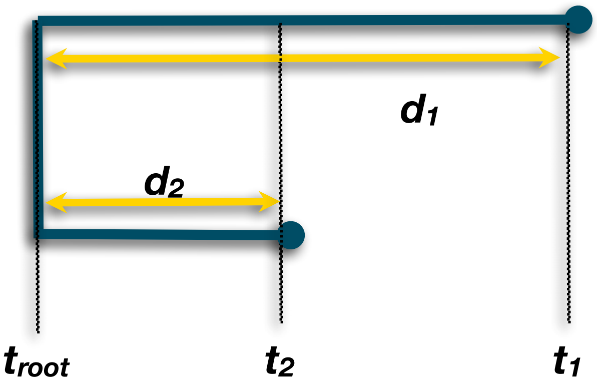


incorporating sampling time: naive method



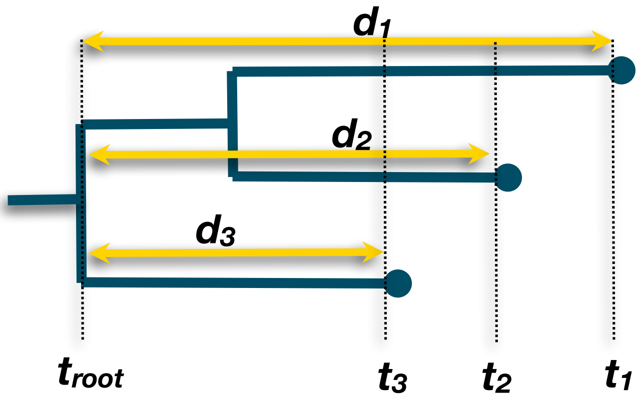


incorporating sampling time: naive method



$$\mu = (d_1 - d_2) / (t_1 - t_2)$$

linear regression



$\mu = d_i / (t_i - t_{root})$

can be rearranged:

$d_i = \mu\ (t_i - t_{root})$

$E[d_i] = \mu\ .\ t_i - \mu\ .\ t_{root}$

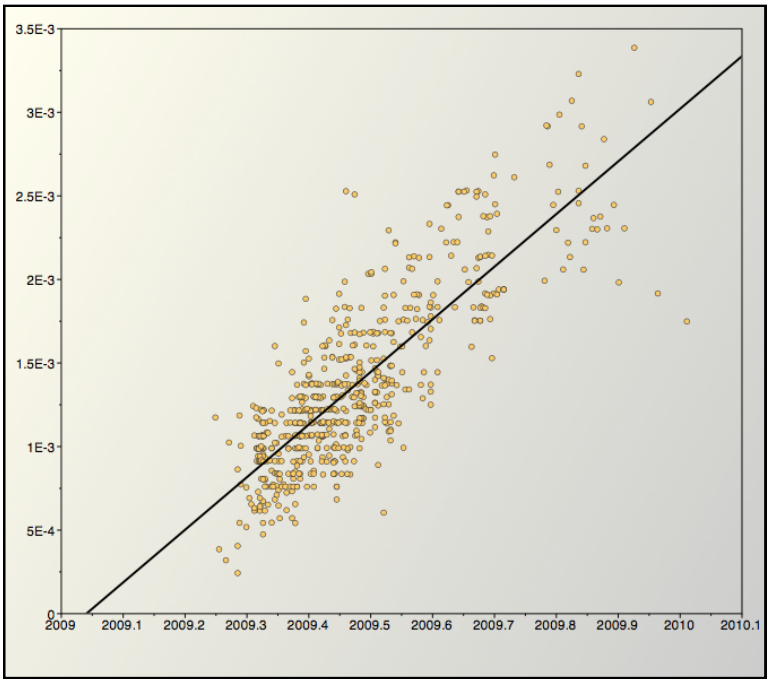
gradient is: μ

y-intercept is: $- \mu\ .\ t_{root}$

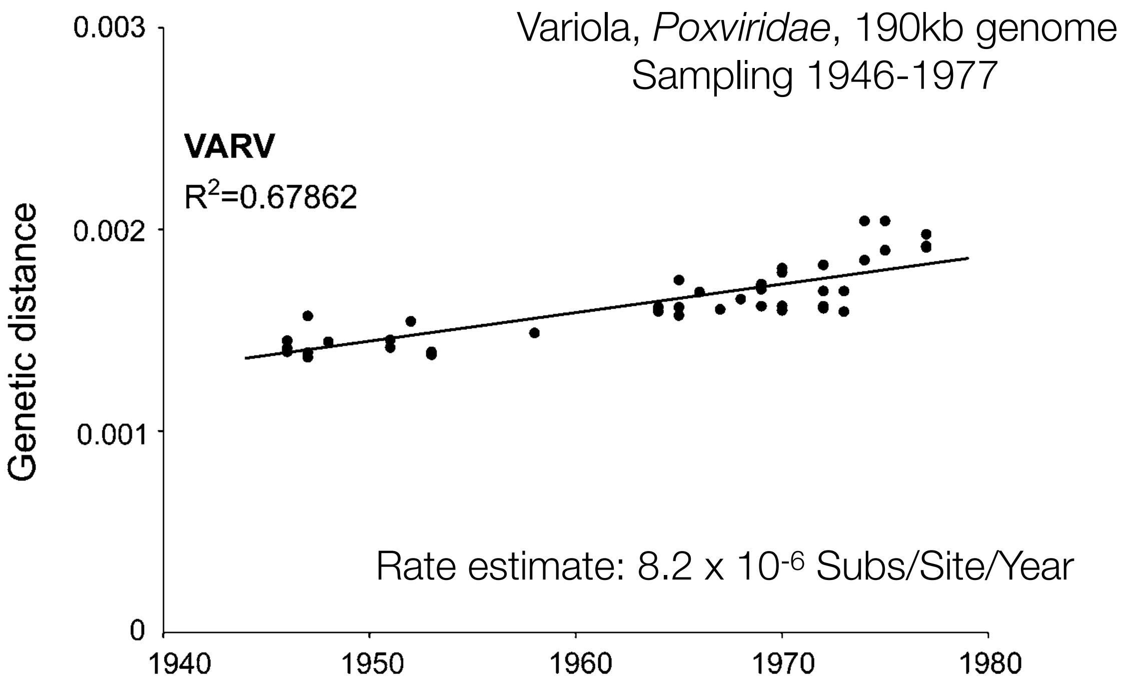
x-intercept is: t_{root}

Estimating the time-scale

- H1N1/09 ‘Swine Flu’
- Rate: 3.14E^{-3}
  mutations/genomic site/year
- tMRCA: 2009.041
  (15-Jan-2009)
- Correlation: 0.83
- R^2: 0.69



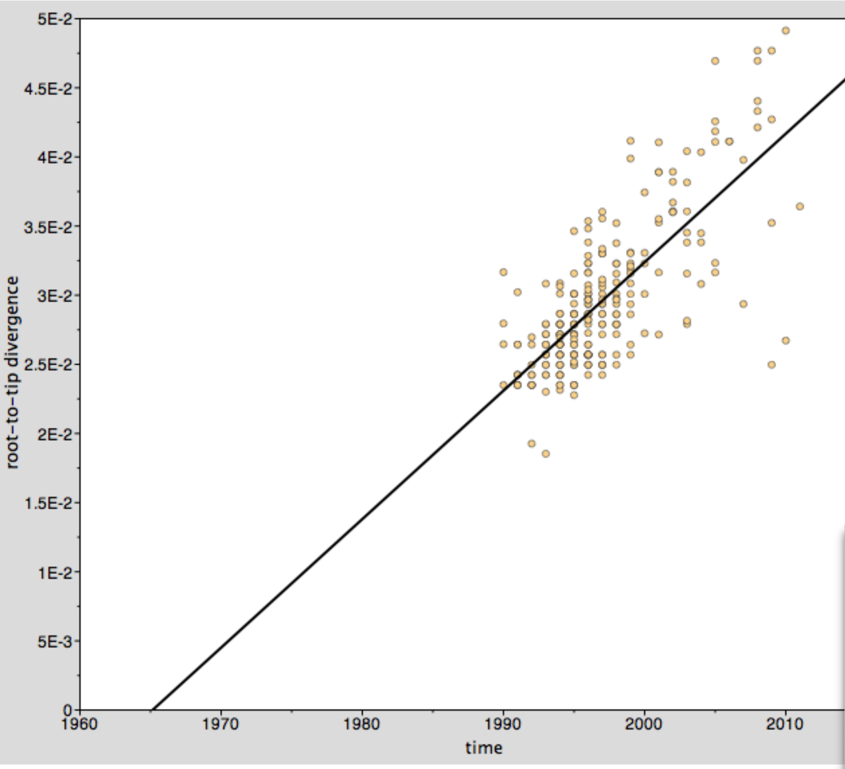
A DNA virus (smallpox)



Variola, *Poxviridae*, 190kb genome
Sampling 1946-1977

Rate estimate: 8.2 x 10^{-6} Subs/Site/Year

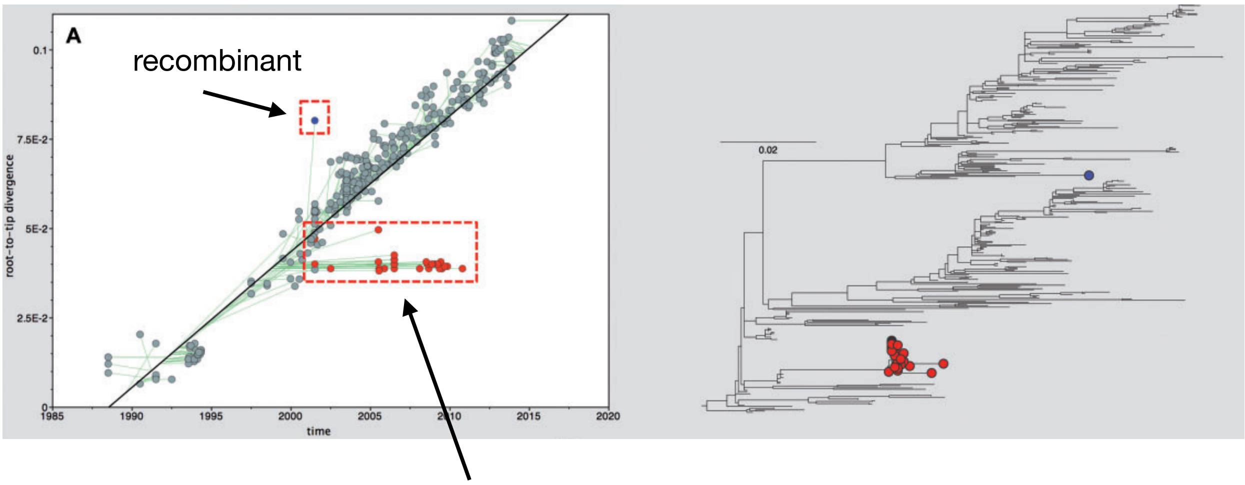
Salmonella Typhimurium



Diagnostic tool

- divergence accumulation
- outliers

TempEst

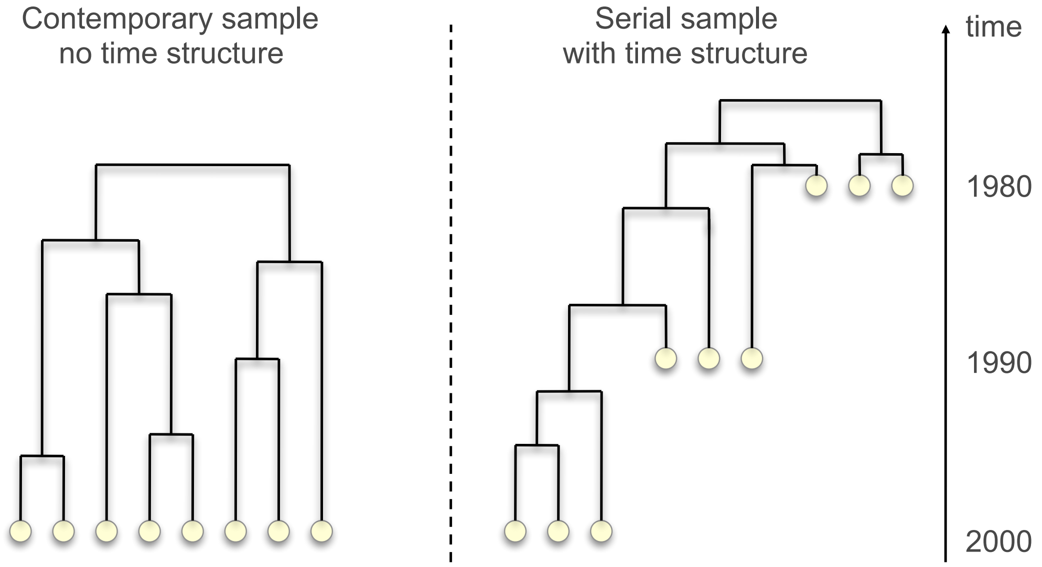


recombinant

vaccine lineage

‣ Rambaut A. et al. (2016) *Virus Evolution*, **2(1)**, vew07.

Time structure via tip calibration



Contemporary sample
no time structure

Serial sample
with time structure

time

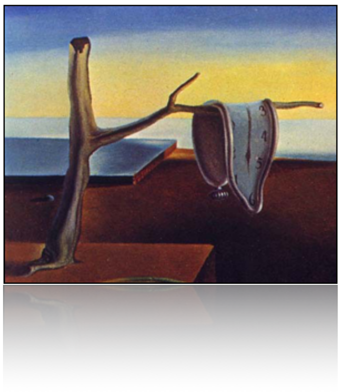
1980

1990

2000

‣ Rambaut A. (2000) *Bioinformatics*, **16**, 395-399.

Relaxing the molecular clock



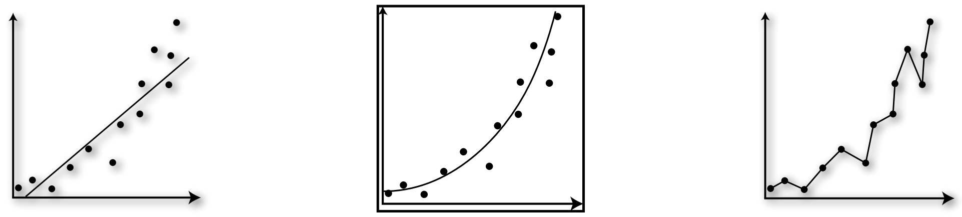
Clock versus non-clock

- unconstrained (unrooted) Felsenstein model:
  Felsenstein (1981) *JME*, **17**: 368 - 376
  - each branch has its own rate independent of all others
  - time and rate are confounded and can only be estimated as a compound parameter (branch lengths)

- strict molecular clock:
  Zuckerkandl & Pauling (1962) in Horizons in Biochemistry, pp. 189–225
  - all lineages evolve at the same rate
  - allows the estimation of the root of the tree and dates of individual nodes

Need for a relaxed molecular clock

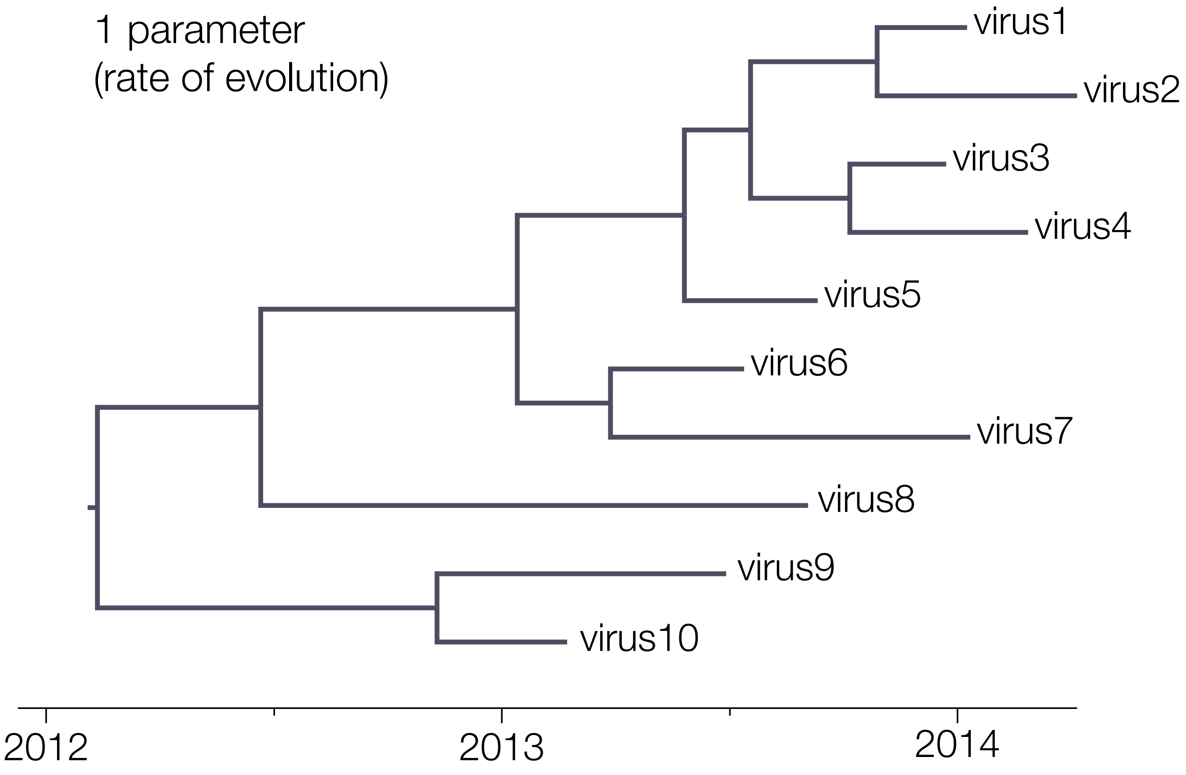
- the unrooted model of phylogeny and the strict molecular clock model are two extremes of a continuum.
- dominate phylogenetic inference
- but both are biologically unrealistic:
  - the real evolutionary process lies between these two extremes
  - model misspecification can produce positively misleading results

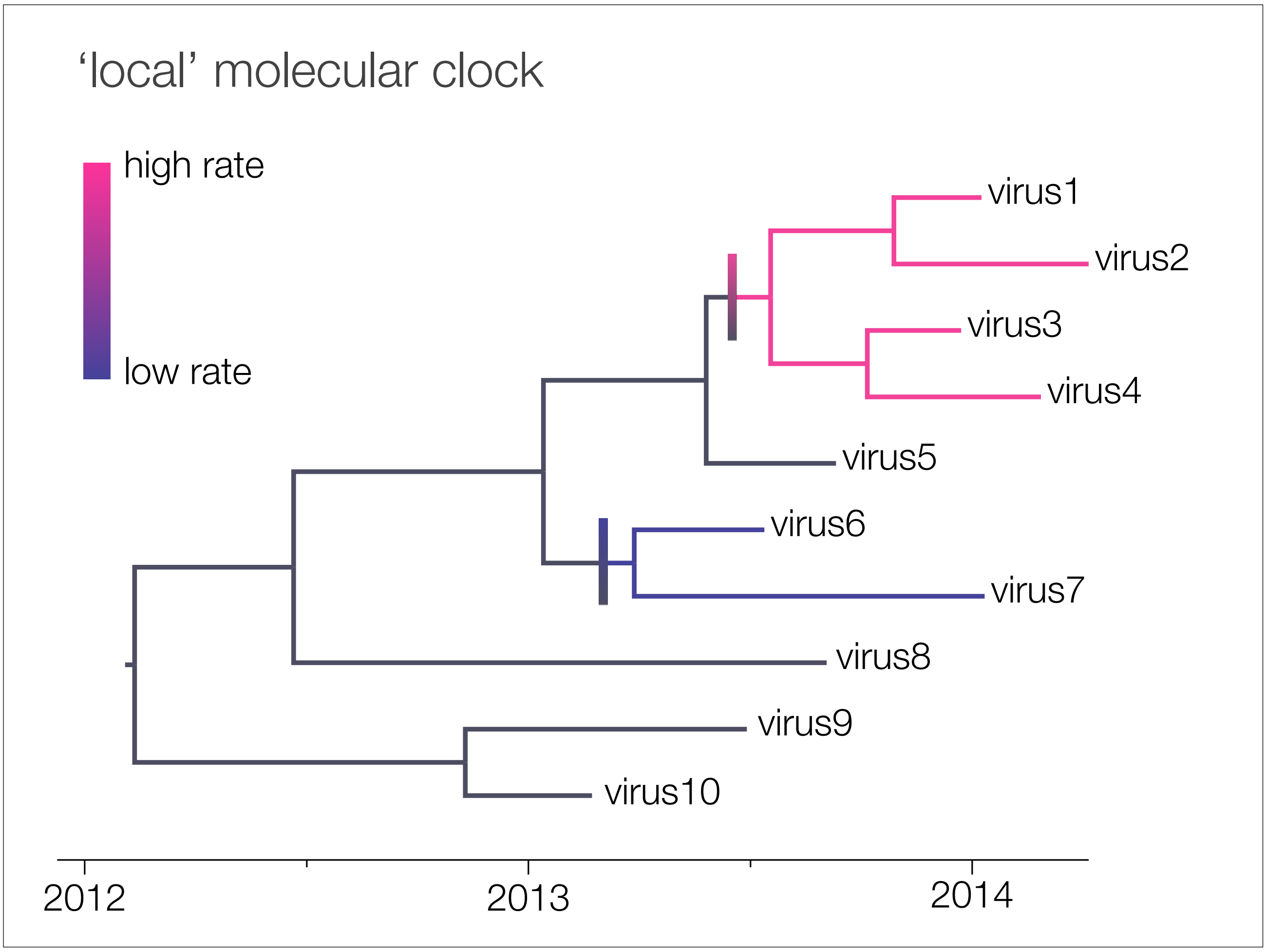


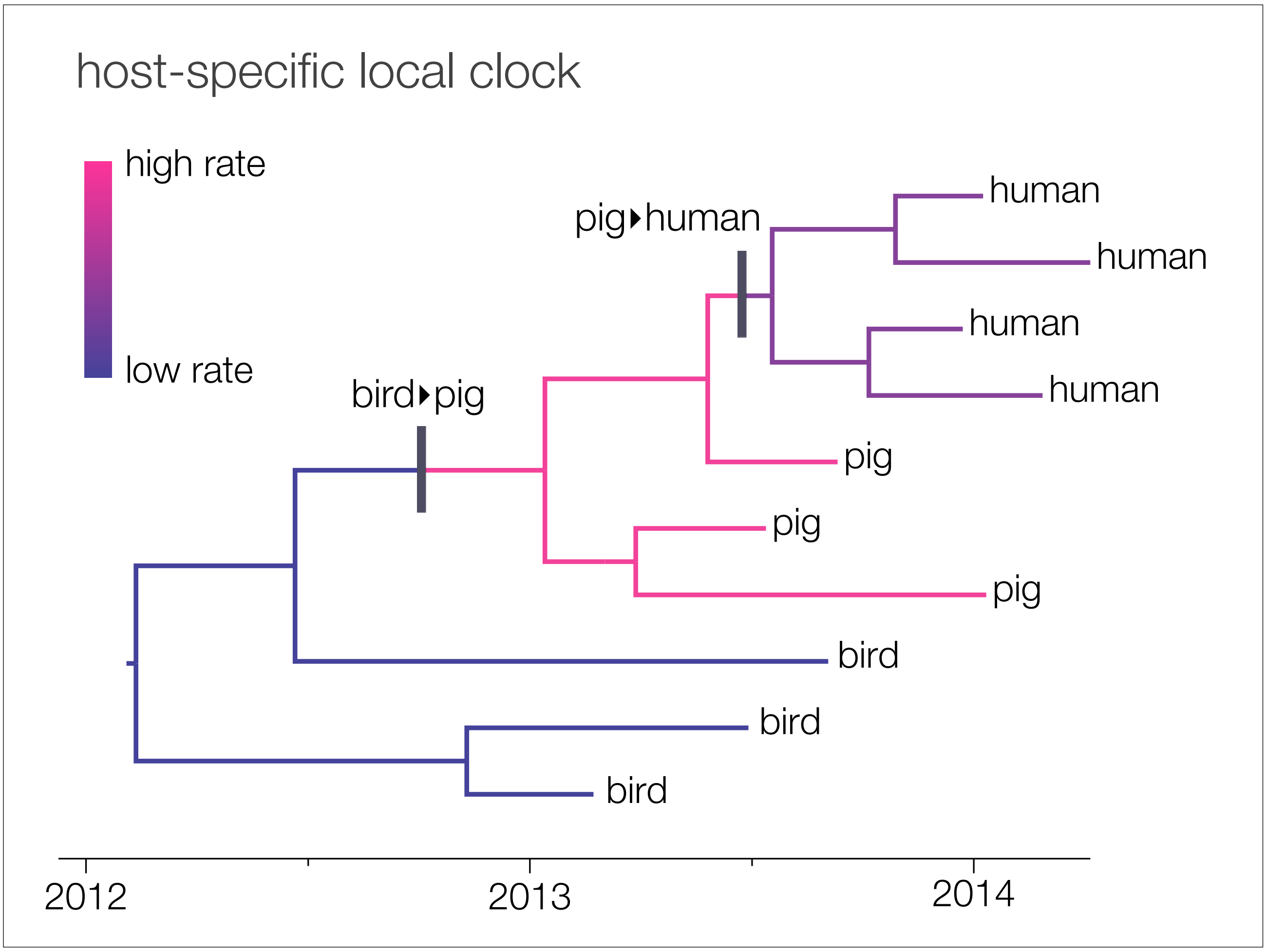
- Pybus (2006) *Genome Biol*. **4**, e151

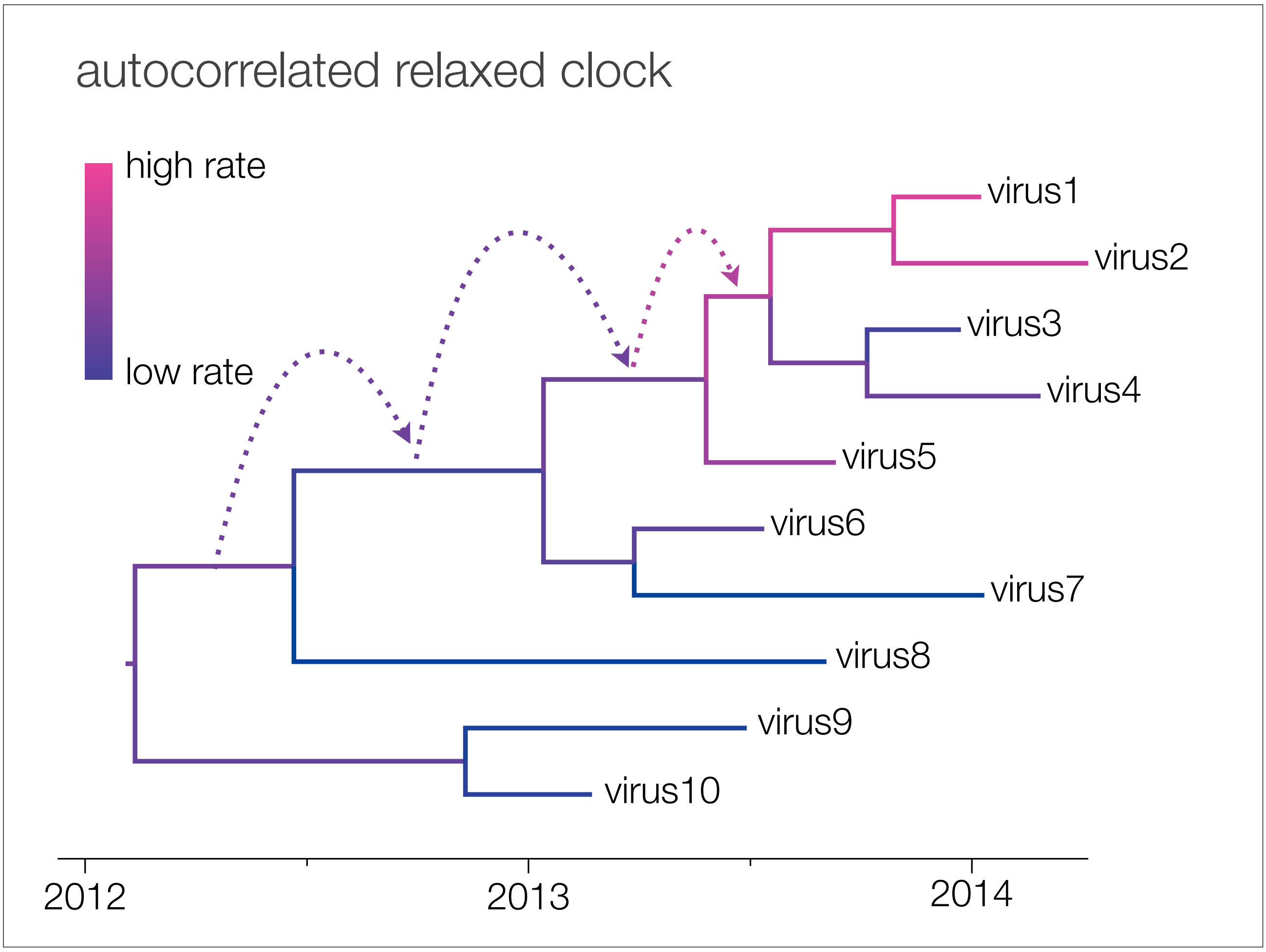
‘strict’ molecular clock

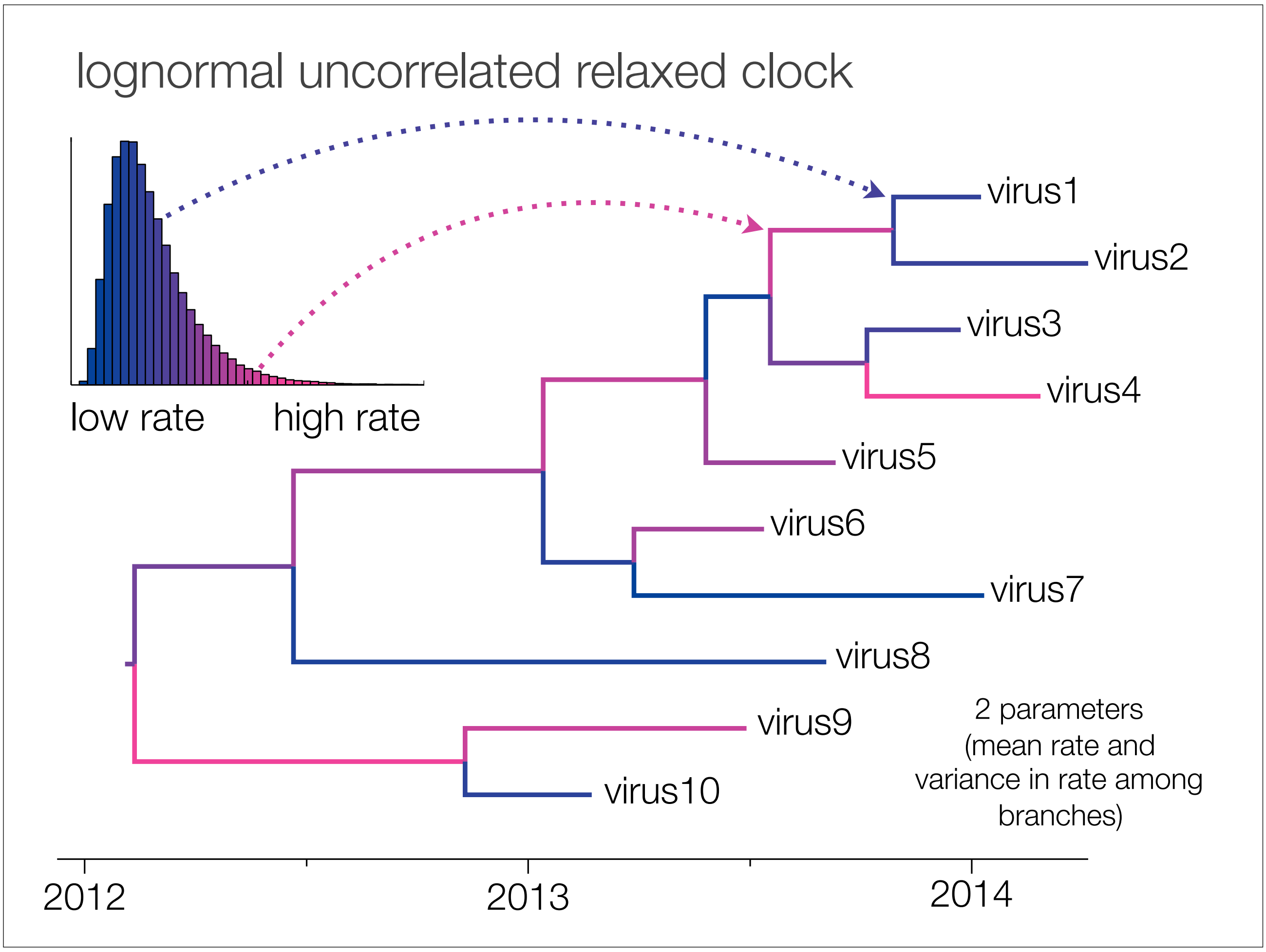
1 parameter
(rate of evolution)




'local' molecular clock
high rate
low rate
virus1
virus2
virus3
virus4
virus5
virus6
virus7
virus8
virus9
virus10
2012
2013
2014

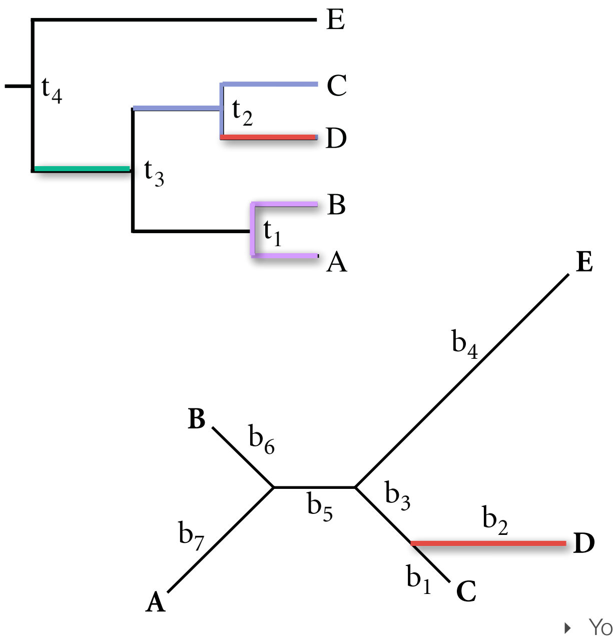

host-specific local clock
high rate
low rate
pig▸human
bird▸pig
human
human
human
human
pig
pig
pig
bird
bird
bird
2012
2013
2014

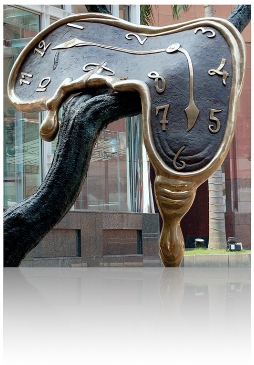




2 parameters
(mean rate and
variance in rate among
branches)

Relaxed clocks: (1) local molecular clocks

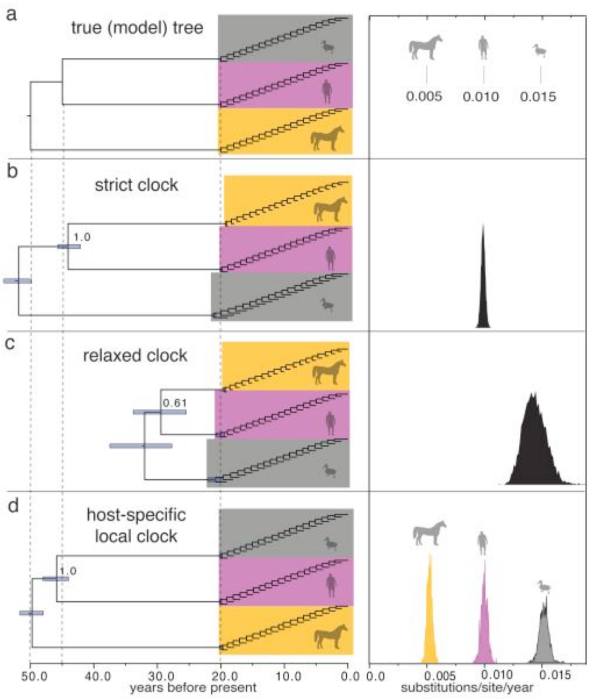


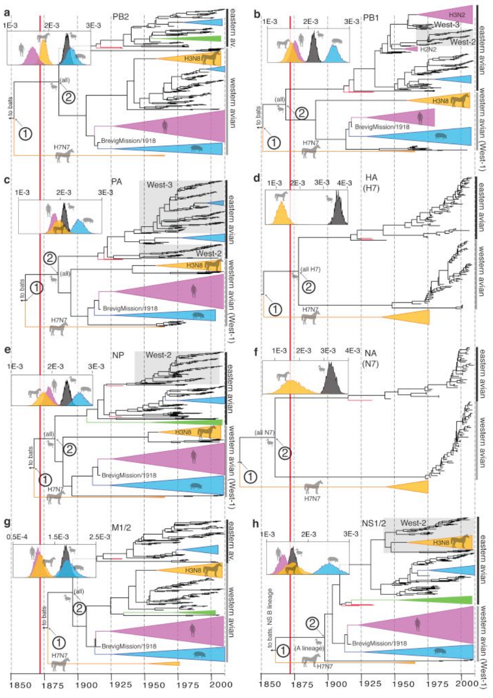


- specify H_0 beforehand
- problem of identifiability

- Yoder and Yang (2000) Mol Biol & Evol **17**: 1081-1090.

Bayesian local clocks

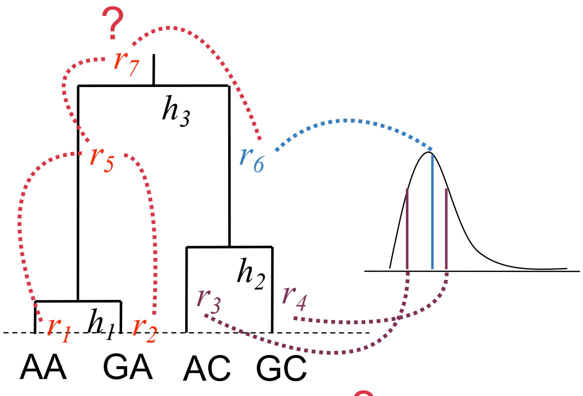




Worobey et al., Nature, 2014; 508(7495): 254–257

Autocorrelated relaxed clocks

- rates for each branch are drawn from a distribution centred on the rate of the ancestor
  - but what is the rate at the root?
  - A prior degree of autocorrelation?
  - (not currently possible to do phylogenetic inference)

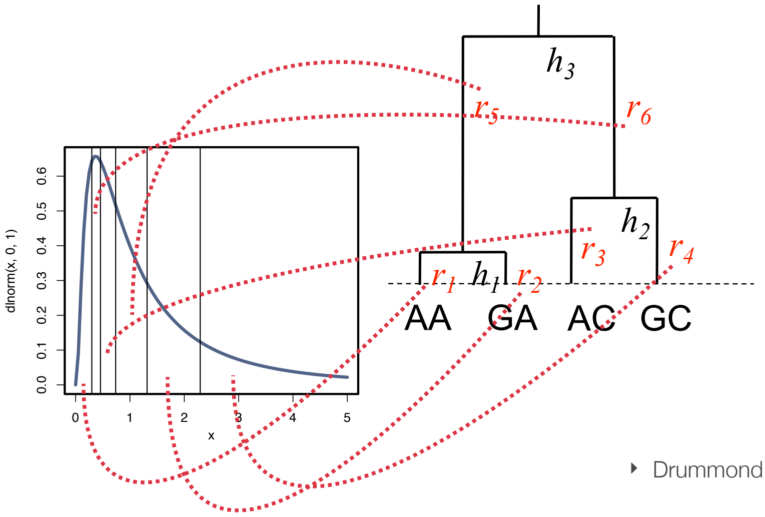


$$r_i \sim LogNormal(r_{A(i)}, \overset{?}{\sigma}^2 \Delta t_i)$$

- e.g., Thorne JL, Kishino H, Painter IS (1998) Mol Biol & Evol **15**: 1647-1657.

Uncorrelated relaxed clocks

- rates for each branch are drawn independently from an identical distribution:



$$r \sim Exp(\lambda)$$

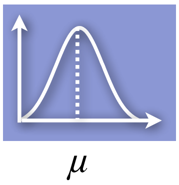
$$r \sim LogNormal(\mu,\sigma^2)$$

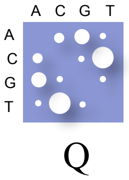
$$r \sim Gamma(\alpha,\beta)$$

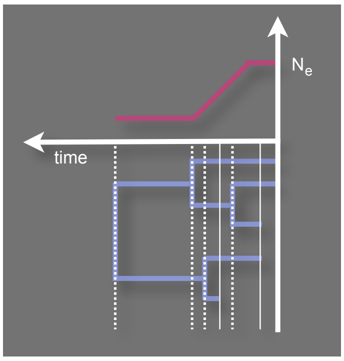
- Drummond et al. (2006) Plos Biology **4**: e88.

Bayesian evolutionary analysis sampling trees

- Given sequence data that is temporally spaced estimate true values of:
  - substitution parameters (μ and Q)
  - ancestral genealogy ($g = E_g, t_Y$)
    - tree topology
    - dates of divergence
  - population history (θ)







- Bayesian inference

$$P(g,\mu,\theta,Q|D) = \frac{1}{Z} \Pr\{D|g,\mu,Q\} f_g(g|\theta) \underline{f_\mu(\mu) f_\theta(\theta) f_Q(Q)}$$

“relaxed phylogenetics and dating with confidence”

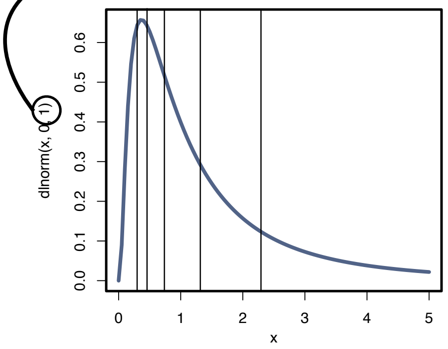
$$t = \{ t_1, t_2, \dots, t_{2n-1} \}$$

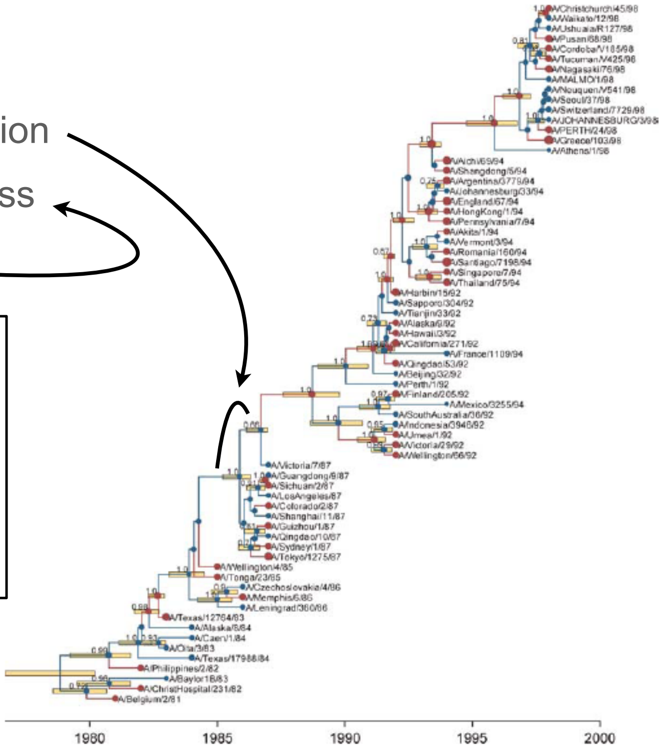
$$R = \{ r_1, r_2, \dots, r_{2n-1} \}$$

$$f(R|g) = f(R) = \prod_{i=1} \lambda e^{-\lambda r_i}$$

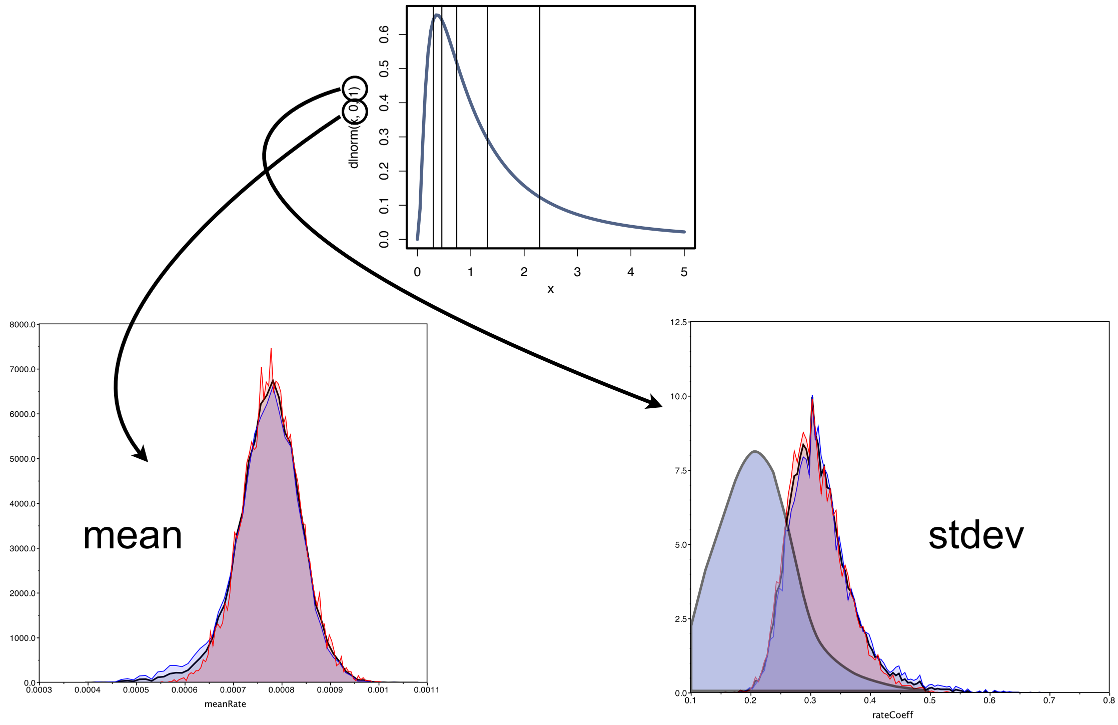
Uncorrelated relaxed clocks: example

- Phylogenetic inference
- measuring autocorrelation
- measuring clock-likeness





Evaluating clock-like behaviour?



Model testing using Bayes factors

- A Bayesian alternative to classical hypothesis testing: the Bayes factor (a summary of the evidence provided by the data in favor of one scientific theory, represented by a statistical model, as opposed to another; Kass & Raftery, 1995).

- Bayes factor $$B_{01} = \frac{p(Y|M_1)}{p(Y|M_0)}$$

- When two models M_0 and M_1 are being compared, one defines the Bayes factor in favor of M_1 over M_0 as the **ratio of their respective marginal likelihoods**

- When there are unknown parameters, the Bayes Factor B_{01} has in a sense the form of a likelihood ratio

Model testing using Bayes factors

- However, the densities are obtained by integrating over parameter space:

$$p(Y|M) = \int_{\theta} p(Y|\theta,M)\ p(\theta|M)\ d\theta$$

- Posterior:

$$p(\theta|Y,M) = \frac{p(Y|\theta,M)\ p(\theta|M)}{p(Y|M)}$$

- So for model fit, the marginal likelihood p(Y|M) or integrated likelihood, i.e. the normalizing constant (cancels out in the calculation of the MH acceptance ratio), is of primary importance, but awfully hard to calculate.

Reminder: MHG MCMC Sampling

The algorithm starts from a random state (θ) and 'proposes' a new state (θ^*)

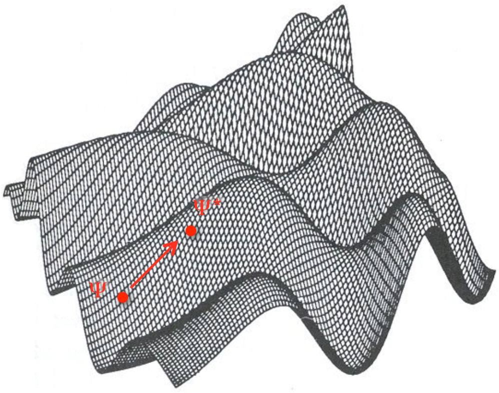
The new state is accepted with probability:

$$R = \min\left(1, \frac{p\ (\theta^*|D)}{p\ (\theta|D)} \times \frac{p\ (\theta|\ \theta^*)}{p\ (\theta^*|\ \theta)}\right)$$

$$= \min\left(1, \frac{p\ (D|\theta^*)\ p\ (\theta^*)/p(D)}{p\ (D|\theta)\ p\ (\theta)/p(D)} \times \frac{f\ (\theta|\ \theta^*)}{f\ (\theta^*|\ \theta)}\right)$$

the two marginal likelihoods cancel out and don't have to be computed !

$$= \min\left(1, \underbrace{\frac{f\ (D|\theta^*)}{f\ (D|\theta)}}_{\text{Likelihood ratio}} \times \underbrace{\frac{f\ (\theta^*)}{f\ (\theta)}}_{\text{Prior ratio}} \times \underbrace{\frac{f\ (\theta|\ \theta^*)}{f\ (\theta^*|\ \theta)}}_{\text{Proposal ratio}}\right)$$



Calculating marginal likelihoods

Methods of general applicability:

- the posterior arithmetic mean estimator (pAME; Aitkin, 1991)
- the arithmetic mean estimator (AME/ILP; but a misnomer)
- the importance sampling estimators, and particularly the harmonic mean estimator (HME) (Newton and Raftery, 1994)
- the stabilized harmonic mean estimator (sHME) (Redelings and Suchard, 2005)

No additional analysis required

- path sampling (Gelman, 1998; Ogata, 1989), applied in phylogenetics (Lartillot and Philippe, 2006)
- stepping-stone sampling (Xie et al., 2011)
- generalised stepping-stone sampling (Fan et al., 2011; Baele et al., 2016)

Additional analysis required

Calculating marginal likelihoods

Methods of general applicability:

- ~~the posterior arithmetic mean estimator (pAME; Aitkin, 1991)~~
- ~~the arithmetic mean estimator (AME/ILP; but a misnomer)~~
- ~~the importance sampling estimators, and particularly the harmonic mean estimator (HME) (Newton and Raftery, 1994)~~
- ~~the stabilized harmonic mean estimator (sHME) (Redelings and Suchard, 2005)~~

~~No additional analysis required~~

- path sampling (Gelman, 1998; Ogata, 1989), applied in phylogenetics (Lartillot and Philippe, 2006)
- stepping-stone sampling (Xie et al., 2011)
- generalised stepping-stone sampling (Fan et al., 2011; Baele et al., 2016)

Additional analysis required

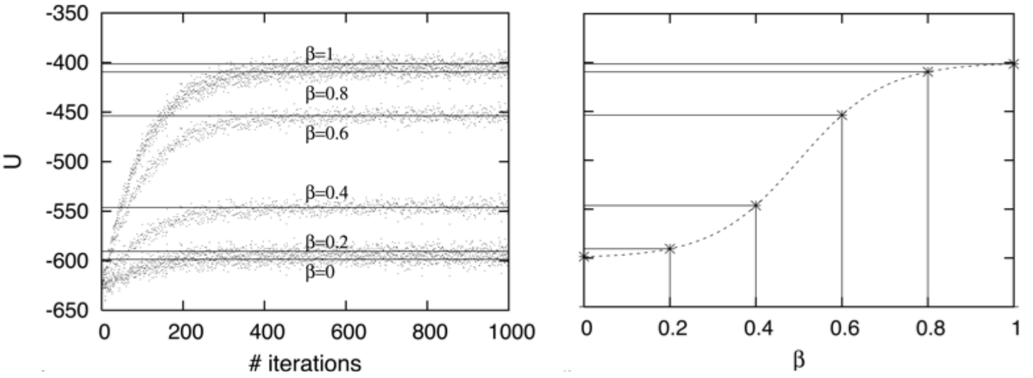
path sampling and stepping-stone sampling

- requires samples from a series of power posteriors, along a path between prior and posterior:

$q_\beta(\theta) = p(Y \mid \theta,M)^\beta p(\theta \mid M)$

reduces to the posterior when $\beta = 1$

reduces to the prior when $\beta = 0$



path sampling and stepping-stone sampling

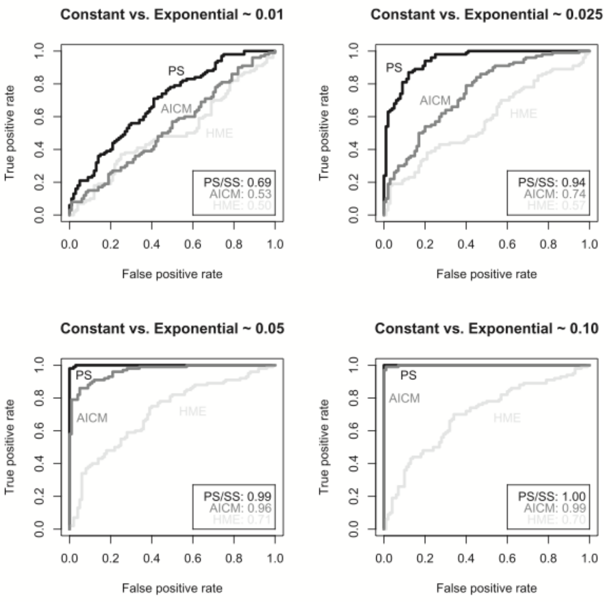


FIG. 2. Evaluation of log BF estimates using PS (SS yields an undistinguishable plot), AICM, and the HME to compare model fit, with four pairwise comparisons being shown: a constant population size versus an exponential population size with growth rates of 0.01, 0.025, 0.05, and 0.10. An increasingly strong discriminatory behavior (low false positive rates and high true positive rates) can be seen for PS (and SS) up to a growth rate of 0.10, whereas the HME retains questionable performance. AICM performance lies in between that of the HME and PS/SS. Color-coded area under the curve values are given at the bottom right of each plot.

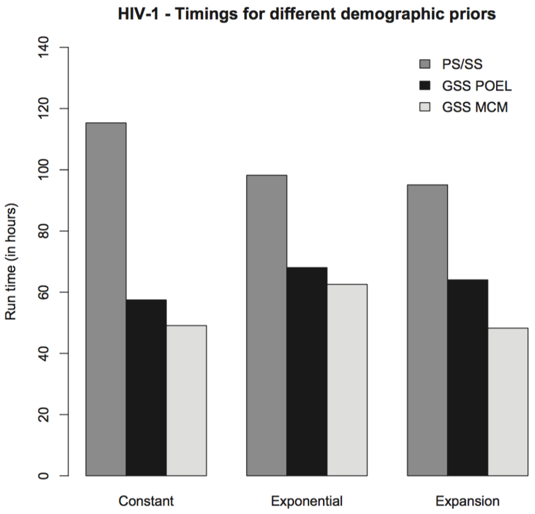
Generalised stepping-stone sampling

requires samples from a series of power posteriors, along a path between reference/working distribution and posterior:

$$q_\beta(\theta) = [p(Y \mid \theta, M)p(\theta \mid M)]^\beta p_0(\theta \mid M)^{1-\beta}$$

- reduces to the original SS method if the reference/working distribution is equal to the actual prior
- in practice, samples from the posterior distribution ($\beta = 1$) are used to parameterize the joint reference/working distribution $p_0(\theta|M)$
- we will use kernel density estimation (KDE) to construct reference/working priors for each of the parameters being estimated

GSS: decreased run time



- GSS does not need to explore the prior, which avoids computing the likelihood for highly unlikely parameter values, which may lead to numerical instabilities
- combined with a “shorter” path to be traversed, this leads to a considerable performance increase (dependent on the actual reference/working prior)

Bayesian model testing

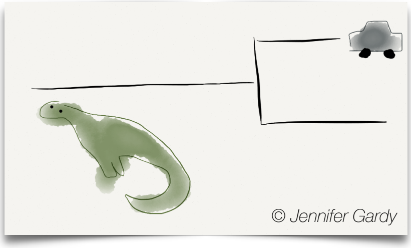
- Don’t compare all possible model combinations (evolutionary model, clock models, coalescent tree prior, …) to one another!
- Test/compare those models if
  - it is part of the hypothesis your testing,
  - or if your hypothesis test is sensitive to the model choice

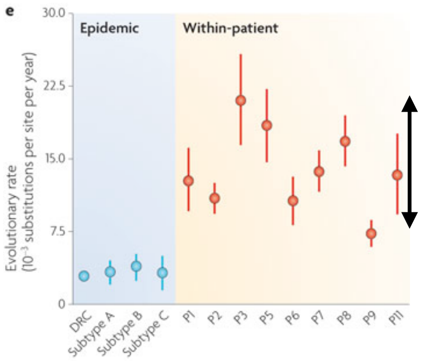
Bayesian model selection vs model averaging

Model selection refers to the problem of using the data to select one model from the list of candidate models

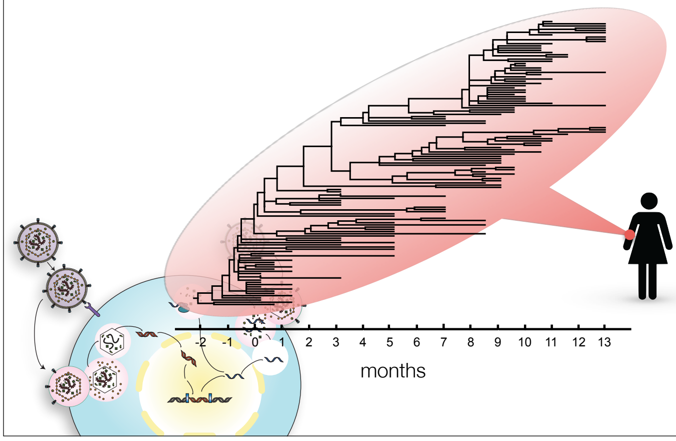
Model averaging refers to the process of estimating some quantity under each model and then averaging the estimates according to how likely each model is.

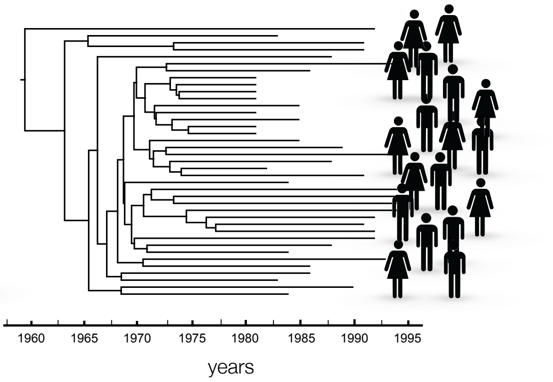
Extensions for testing evolutionary rate hypotheses





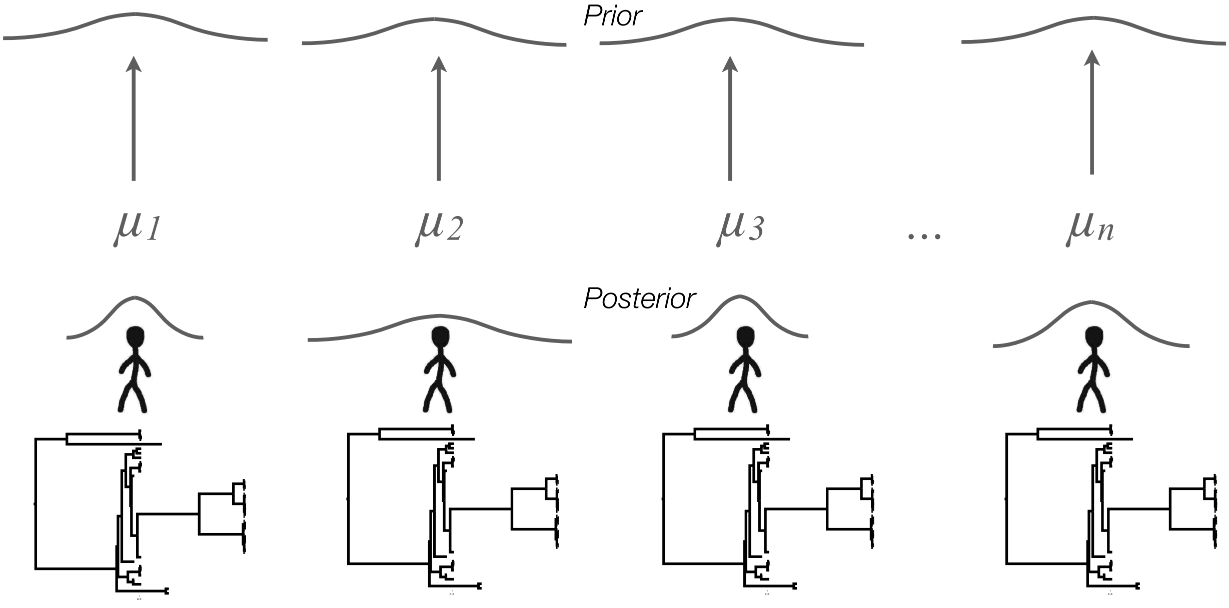
Pybus and Rambaut, NGR, 2009



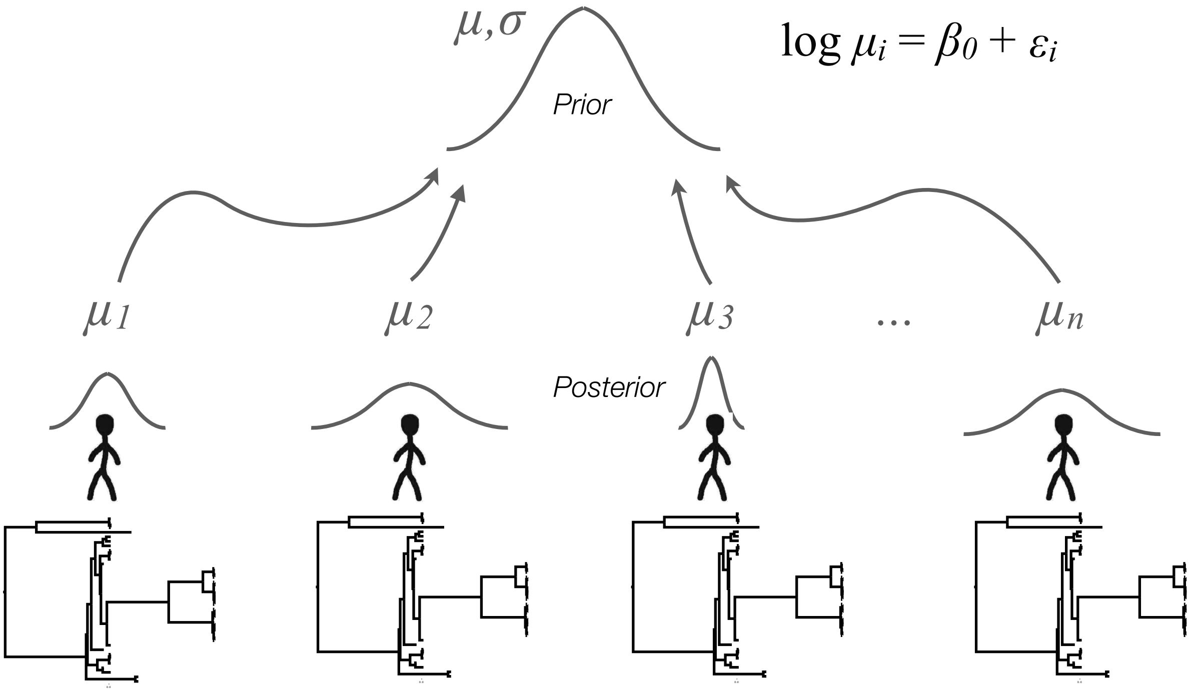


Lemey et al 2006 AIDS Rev

Independent parameter estimation



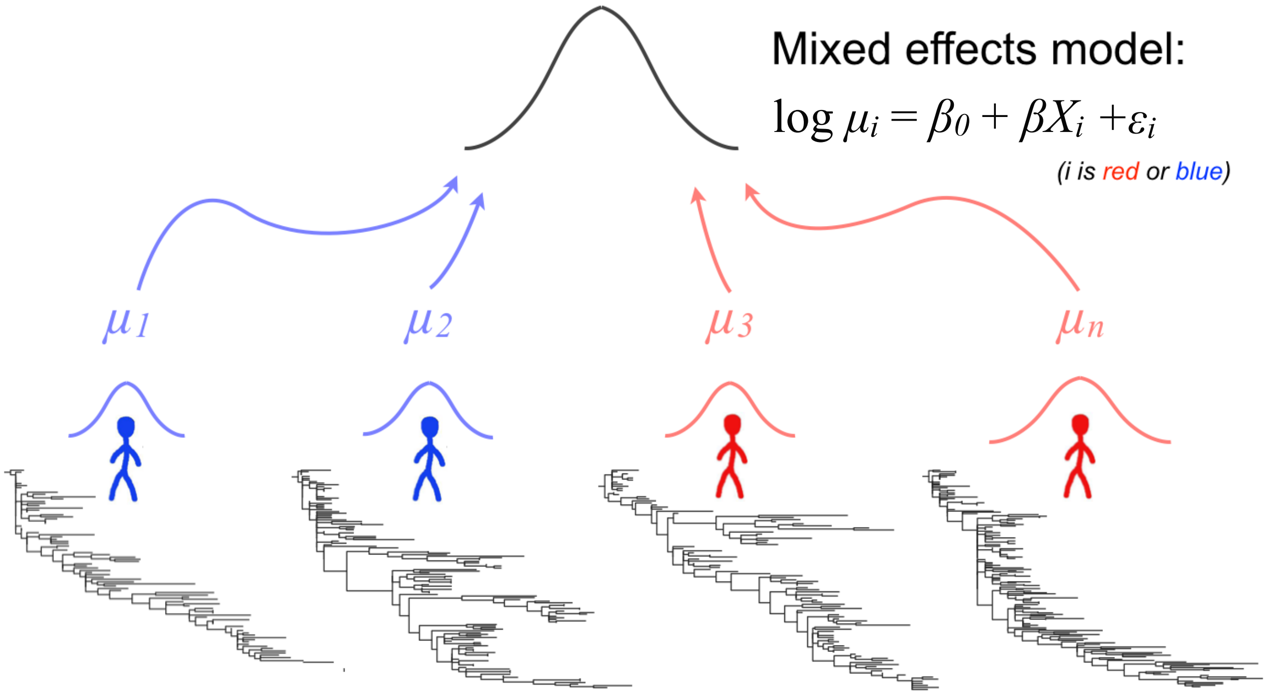
Hierarchical phylogenetic models



$$\log \mu_i = \beta_0 + \varepsilon_i$$

Edo-Matas et al., MBE, 2011

Hierarchical model with fixed effects



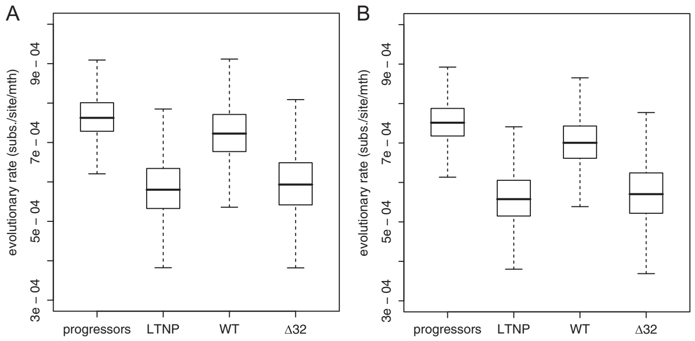
Mixed effects model:

$$\log \mu_i = \beta_0 + \beta X_i + \varepsilon_i$$

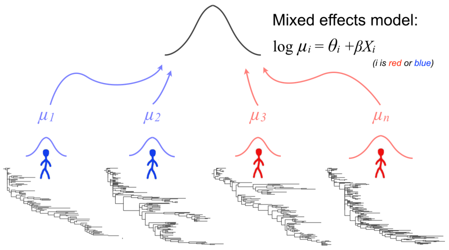
(*i* is *red* or *blue*)

Edo-Matas et al., MBE, 2011

Hierarchical model with fixed effects



$$\log\theta_i = \beta_0 + \delta_{\mathrm{LTNP}}\beta_{\mathrm{LTNP}}\mathrm{LTNP}_i + \delta_{\Delta 32}\beta_{\Delta 32}\Delta 32_i + \varepsilon_i$$



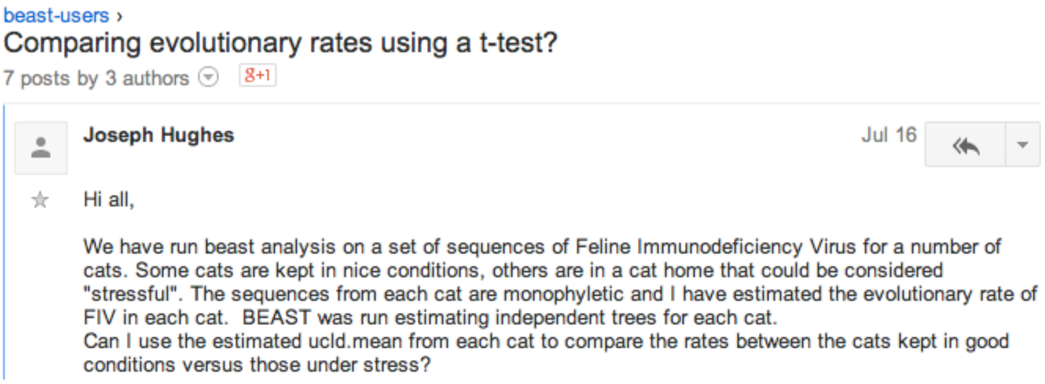
Mixed effects model:

$$\log \mu_i = \theta_i + \beta X_i$$

(*i* is *red* or *blue*)

Evolutionary Parameter	Effect Support/Size	LTNP Effect
Nucleotide substitution rate	Posterior probability $\delta_{\mathrm{effect}} = 1$	0.72
	$\mathrm{BF}_{\mathrm{effect}}$	2.6
	$\beta_{\mathrm{effect}}\|\delta_{\mathrm{effect}} = 1$[a]	−0.275 (−0.524, −0.016)
Codon substitution rate	Posterior probability $\delta_{\mathrm{effect}} = 1$	0.726
	$\mathrm{BF}_{\mathrm{effect}}$	2.6
	$\beta_{\mathrm{effect}}\|\delta_{\mathrm{effect}} = 1$[a]	−0.265 (−0.523,0.019)
d_N/d_S	Posterior probability $\delta_{\mathrm{effect}} = 1$	0.502
	$\mathrm{BF}_{\mathrm{effect}}$	1.0
	$\beta_{\mathrm{effect}}\|\delta_{\mathrm{effect}} = 1$[a]	0.083 (−0.101,0.25)

Edo-Matas et al., MBE, 2011



beast-users ›

Comparing evolutionary rates using a t-test?

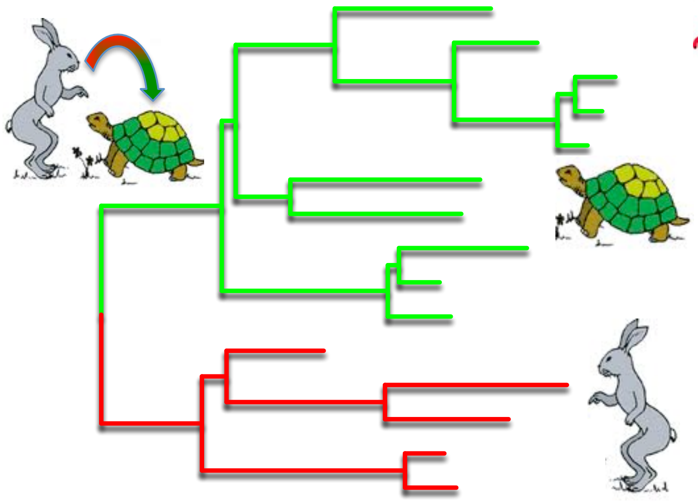
7 posts by 3 authors

Joseph Hughes — Jul 16

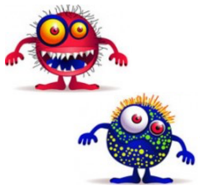
Hi all,

We have run beast analysis on a set of sequences of Feline Immunodeficiency Virus for a number of cats. Some cats are kept in nice conditions, others are in a cat home that could be considered "stressful". The sequences from each cat are monophyletic and I have estimated the evolutionary rate of FIV in each cat. BEAST was run estimating independent trees for each cat.
Can I use the estimated ucld.mean from each cat to compare the rates between the cats kept in good conditions versus those under stress?

What drives the tempo of pathogen evolution?



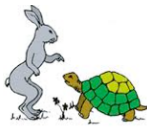
<u>Pathogen factors</u>



Mutation rate

Life cycle/replication dynamics

<u>Host factors</u>

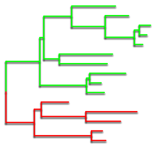


Life history

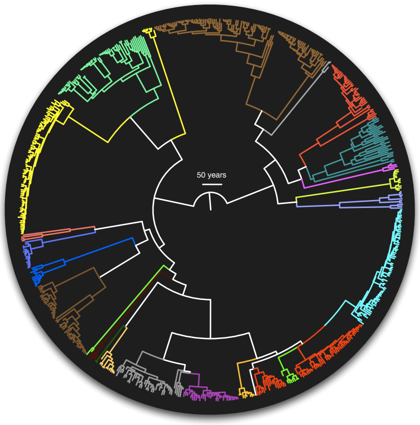
Seasonality

Metabolic rate etc.

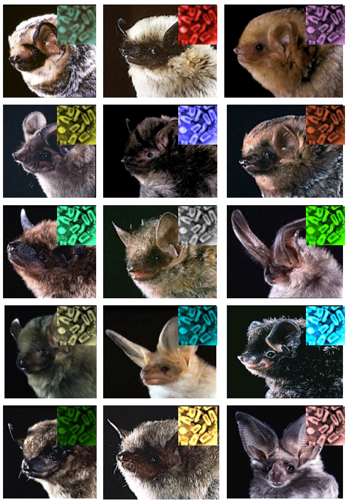
<u>Historical factors</u>



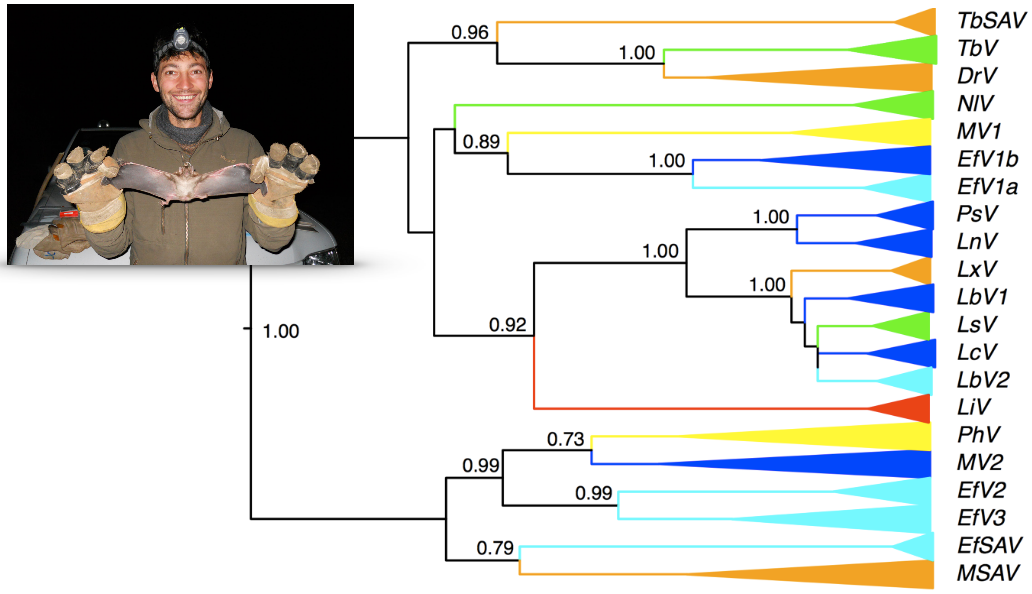
Pathogen phylogeny



Courtesy of D. Streicker



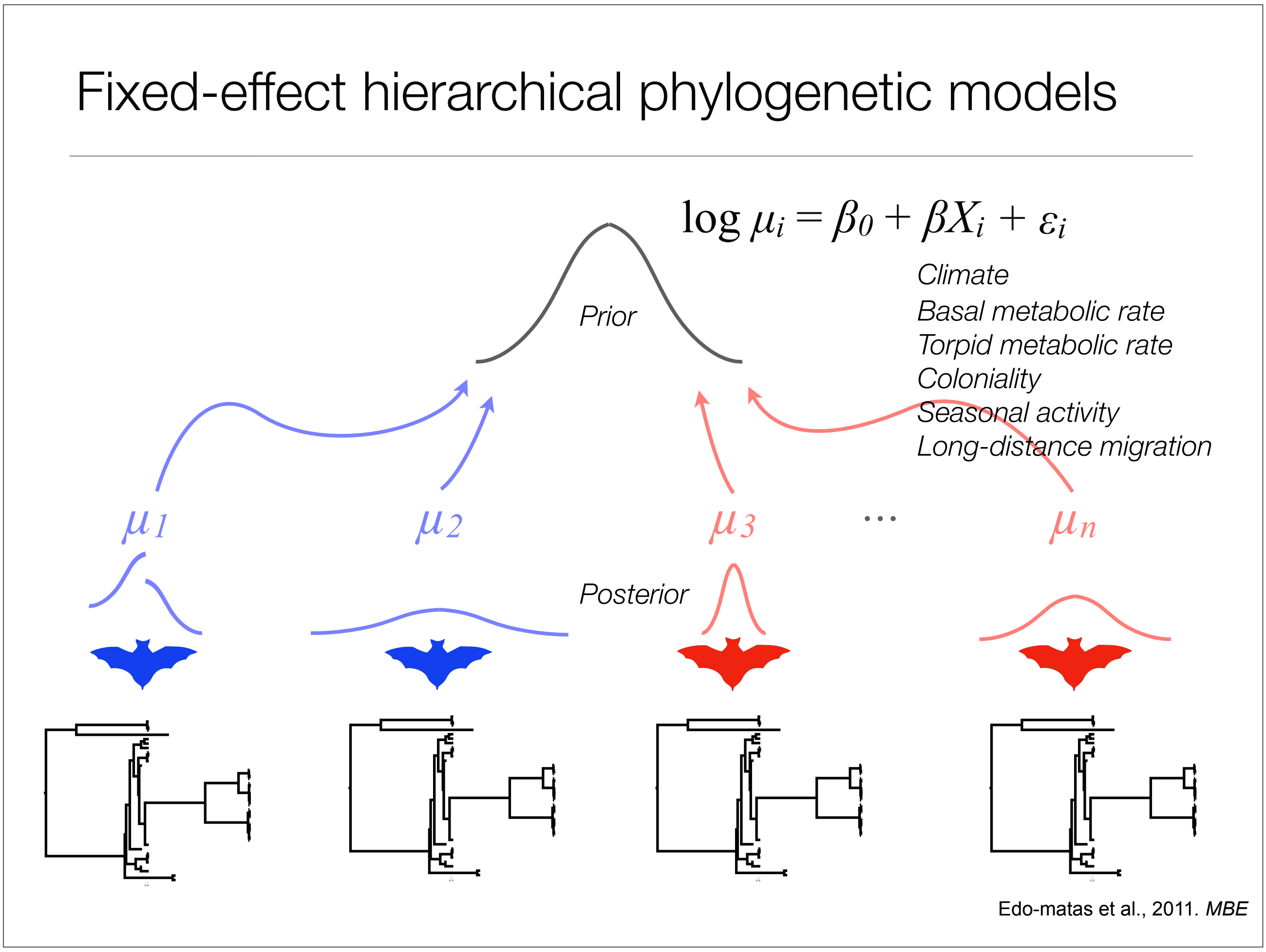
Bat rabies virus evolutionary rates



● 8.79e-5 - 4.22e-4 ○ 4.23e-4 - 7.55e-4 ● 7.56e-4 - 1.09e-3

○ 1.10e-3 - 1.42e-3 ● 1.43e-3 - 1.76e-3 ● 1.77e-3 - 2.09e-3

Streicker et al., 2012. *PLoS Pathogens*

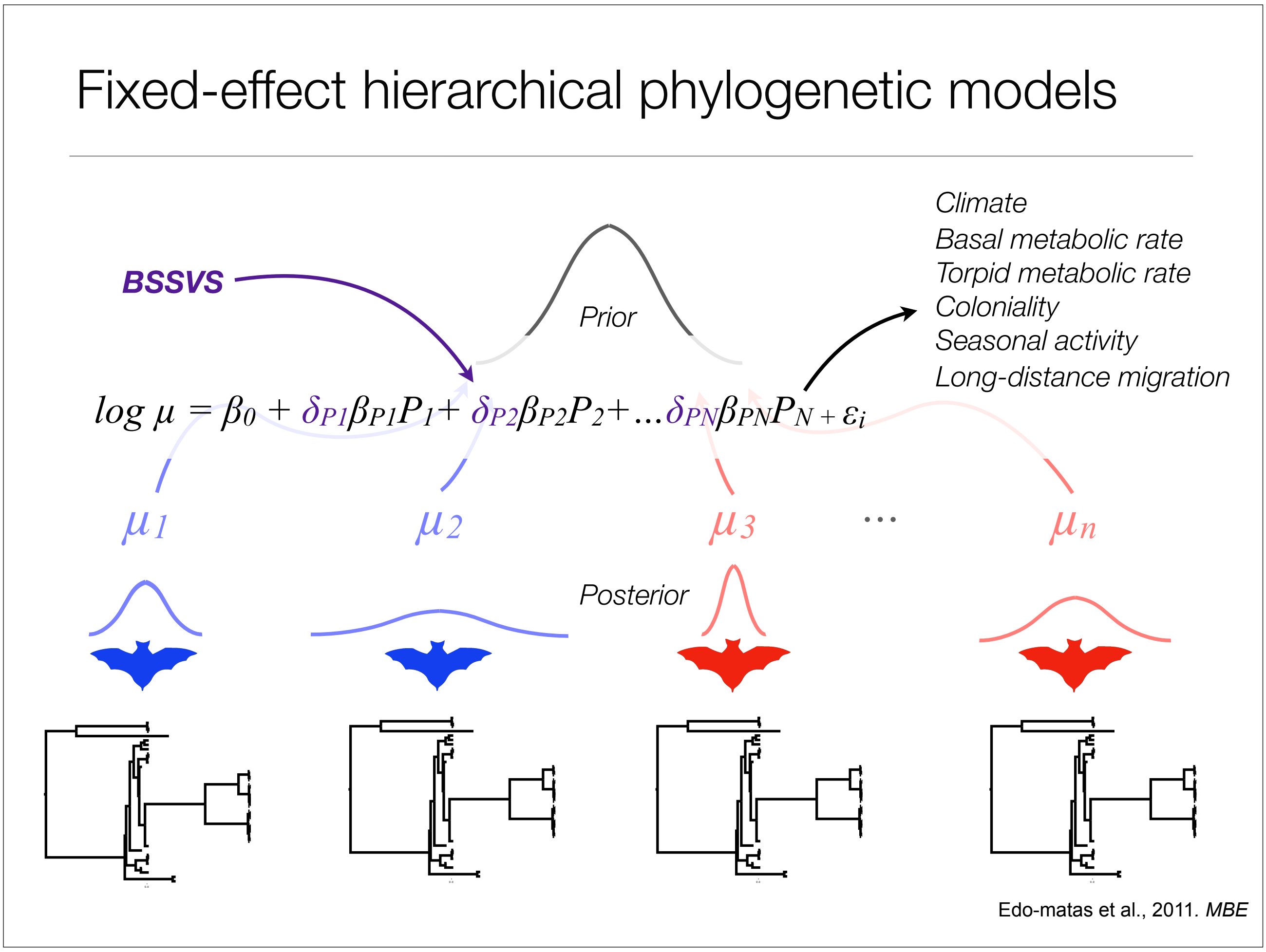


Fixed-effect hierarchical phylogenetic models

$$\log \mu_i = \beta_0 + \beta X_i + \varepsilon_i$$

Climate
Basal metabolic rate
Torpid metabolic rate
Coloniality
Seasonal activity
Long-distance migration

Edo-matas et al., 2011. *MBE*



Fixed-effect hierarchical phylogenetic models

Climate
Basal metabolic rate
Torpid metabolic rate
Coloniality
Seasonal activity
Long-distance migration

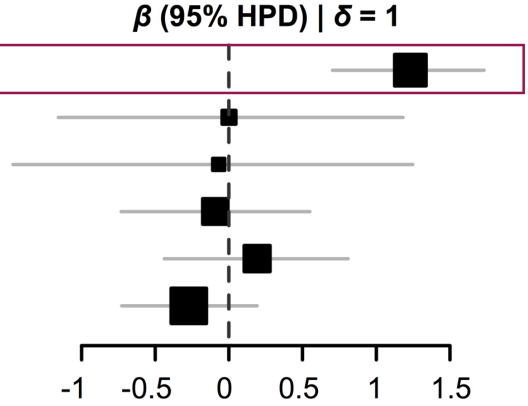
BSSVS

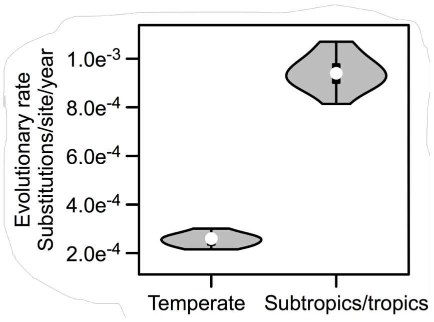
$$\log \mu = \beta_0 + \delta_{P1}\beta_{P1}P_1 + \delta_{P2}\beta_{P2}P_2 + \ldots\delta_{PN}\beta_{PN}P_{N} + \varepsilon_i$$

Edo-matas et al., 2011. *MBE*

Bat rabies virus evolutionary rates

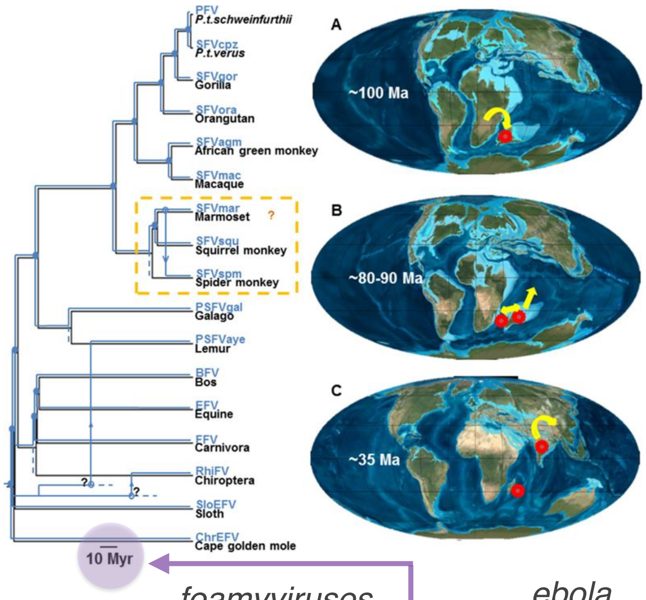
Predictor	Bayes factor
Climate	466.54
Basal metabolic rate	0.82
Torpid metabolic rate	1.00
Coloniality	0.46
Seasonal activity	0.46
Long-distance migration	0.69





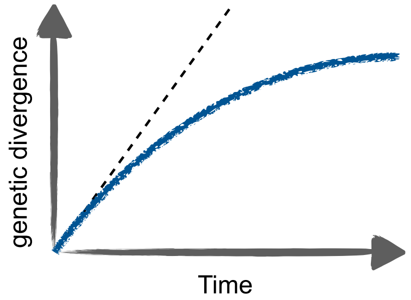
Streicker et al., 2012. *PLoS Pathogens*

Katzourakis et al., Retrovirology, 2014.

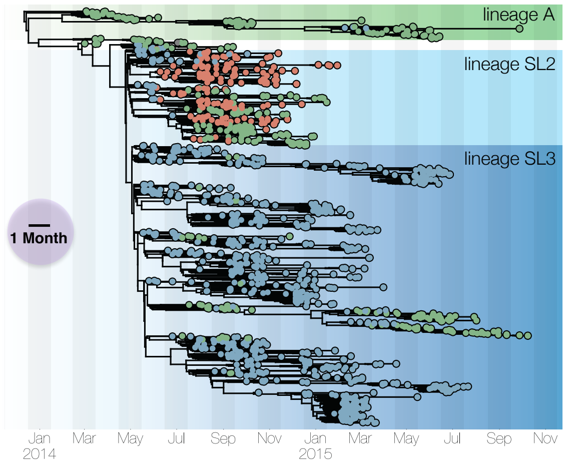


foamyviruses

ebola

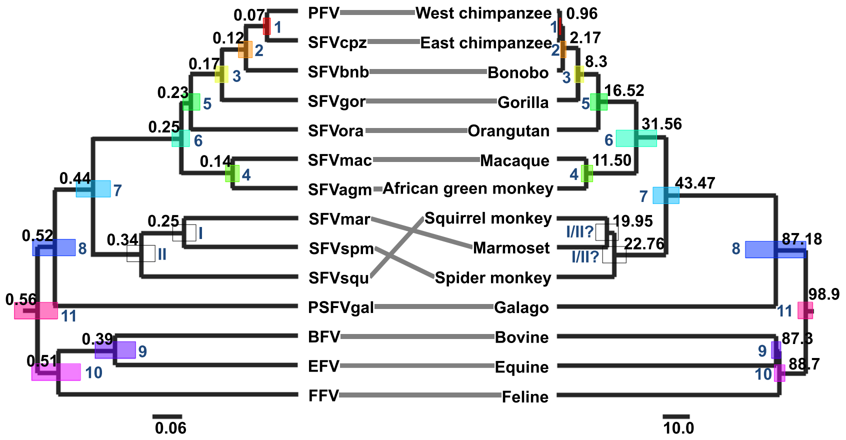


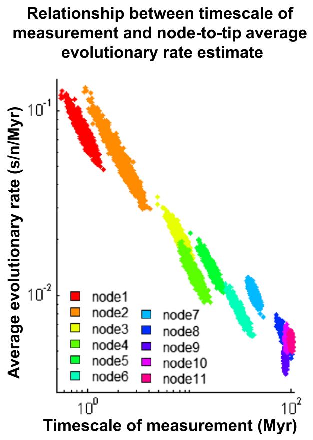
Dudas et al., Nature, 2017.



challenges

Time-dependent evolutionary rates



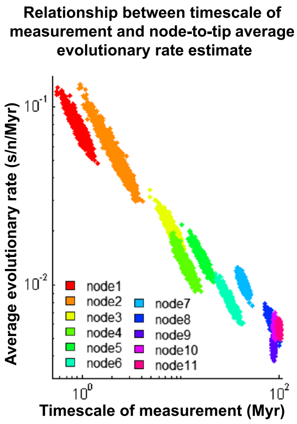


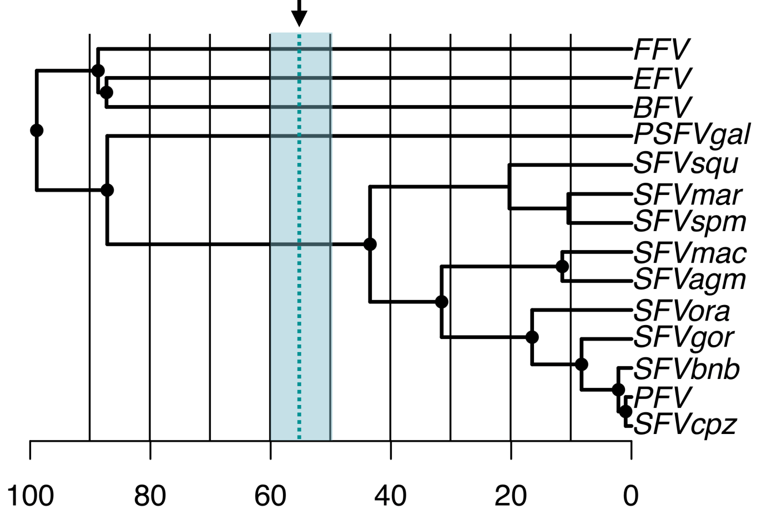
Aiewsakun et al., BMC Evol Biol, 2015.

epoch modelling with TDR

$$\log \mu_i = \beta_0 + \beta_1 X_i$$

$$\log (T_i)$$





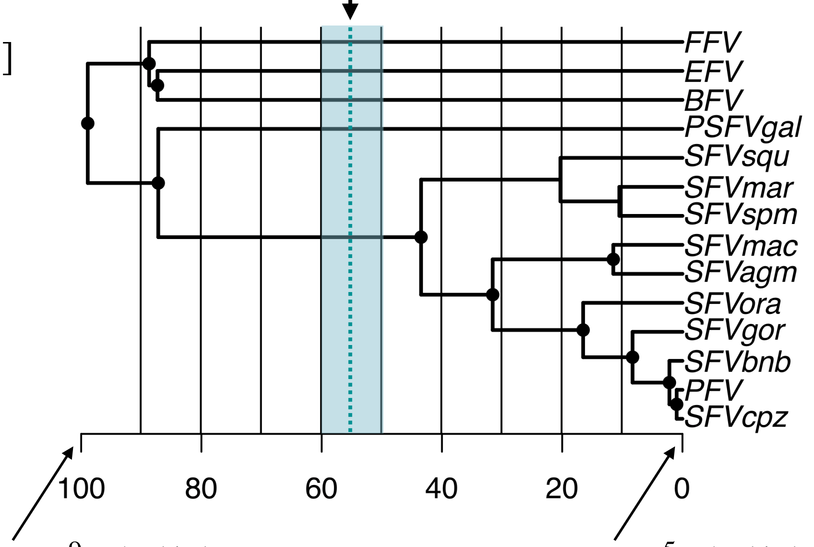
Membrebe et al., 2019. *MBE*

epoch modelling with TDR

$\log \mu_i = \beta_0 + \beta_1 X_i$ $\longrightarrow$ $\log (T_i)$

β_1=-0.539 [-0.570,-0.511]

model	lnL
epoch TDR	-33,667
strict	-34,044



1.77×10^{-9} subst./site/yr

[1.05×10^{-9}, 2.94×10^{-9}]

2.64×10^{-5} subst./site/yr

[2.02×10^{-5}, 3.52×10^{-5}]

Membrebe et al., 2019. *MBE*

conclusions

- molecular clocks: rate constancy assumption and tick rate calibration
- unconstrained <-> strict molecular clock
- relaxed clocks
- model testing: use wisely
- hypotheses -> incorporate them into your model

[redacted]