



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 Lemey¹ and Marc Suchard²

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 19-21, 2017

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

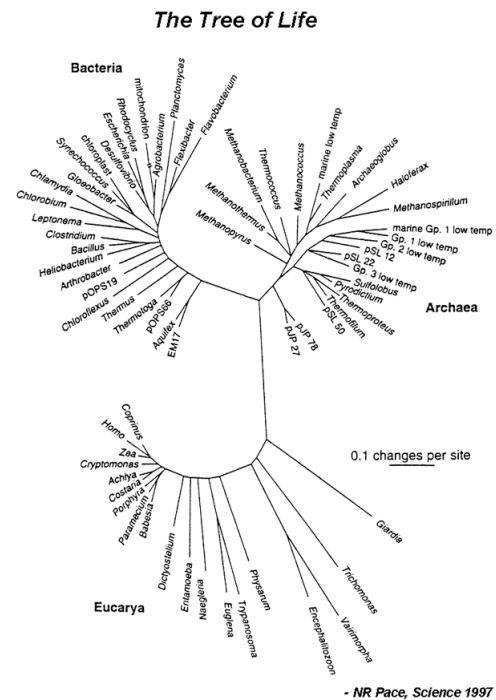
POPULATION GENETICS



EPIDEMIOLOGY

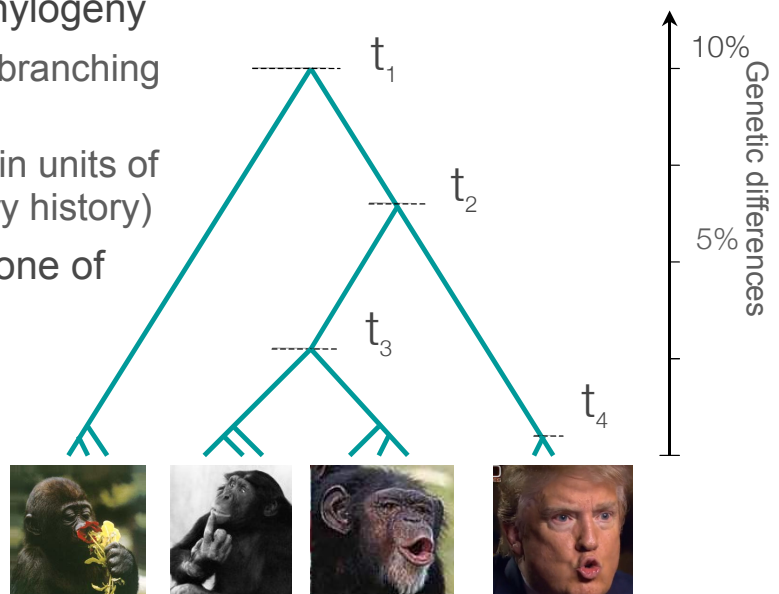
Molecular phylogenies

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



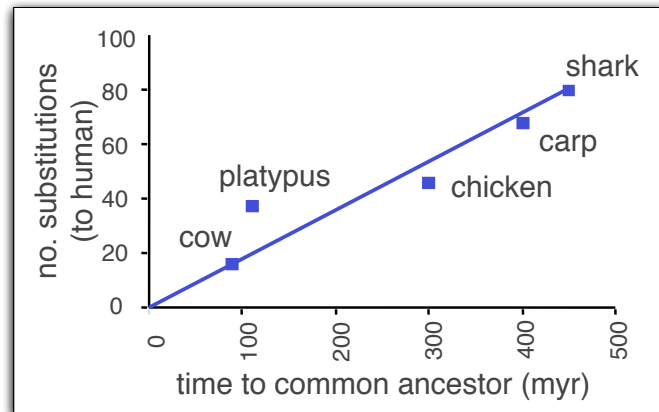
Molecular phylogenies

- 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

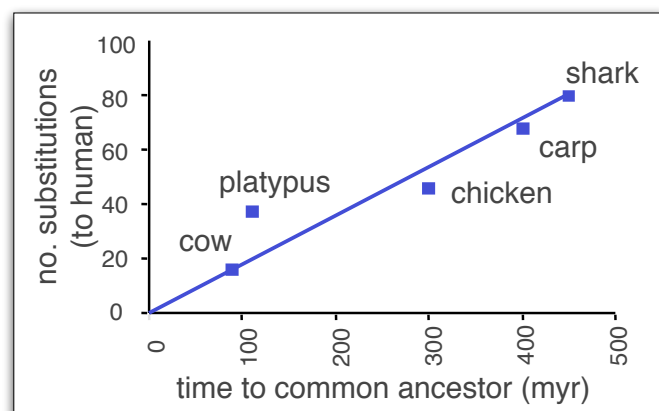
- to obtain a timed phylogeny, the evolutionary model must assume a relationship between the accumulation of genetic diversity and time



- Zuckerkindl 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

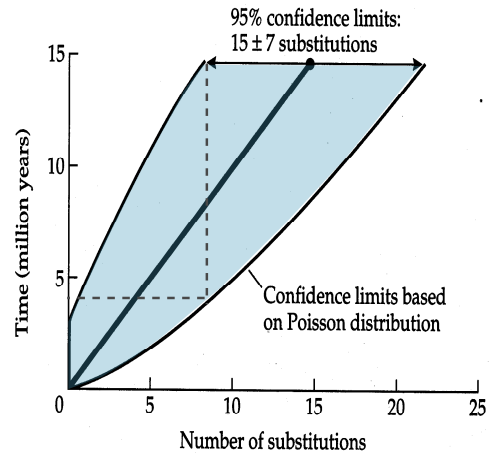
- the *molecular clock* is particularly striking when compared to the obvious differences in rates of morphological evolution...



The molecular clock is not a metronome

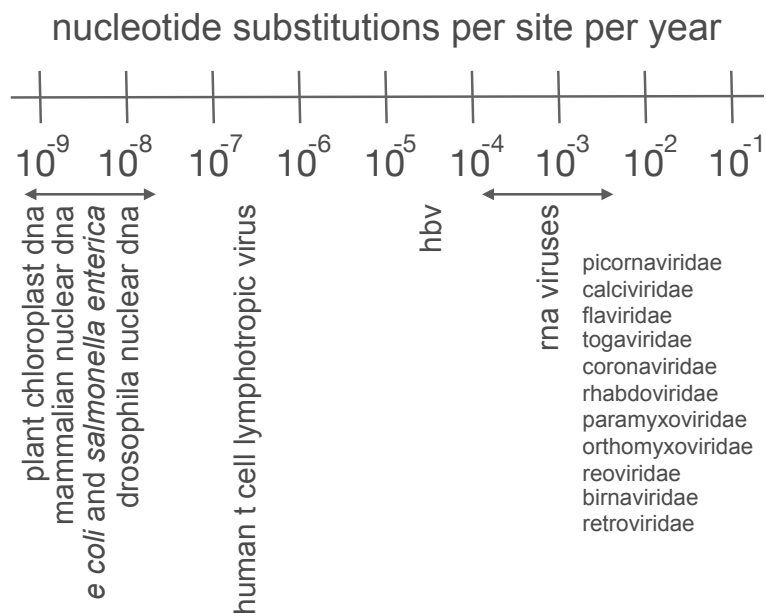
- if mutation every MY with Poisson variance

- ▶ 95% of the lineages 15MY old have 8-22 substitutions
- ▶ 8 substitutions also could be < 5 MY old



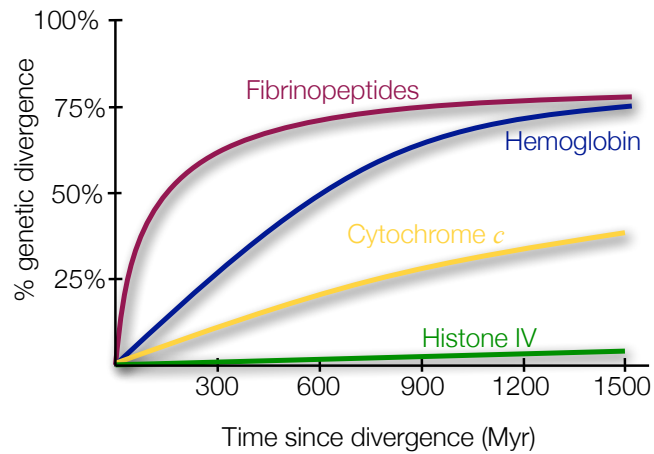
▶ Molecular Systematics, p532.

And there is no global molecular clock



And there is no global molecular clock

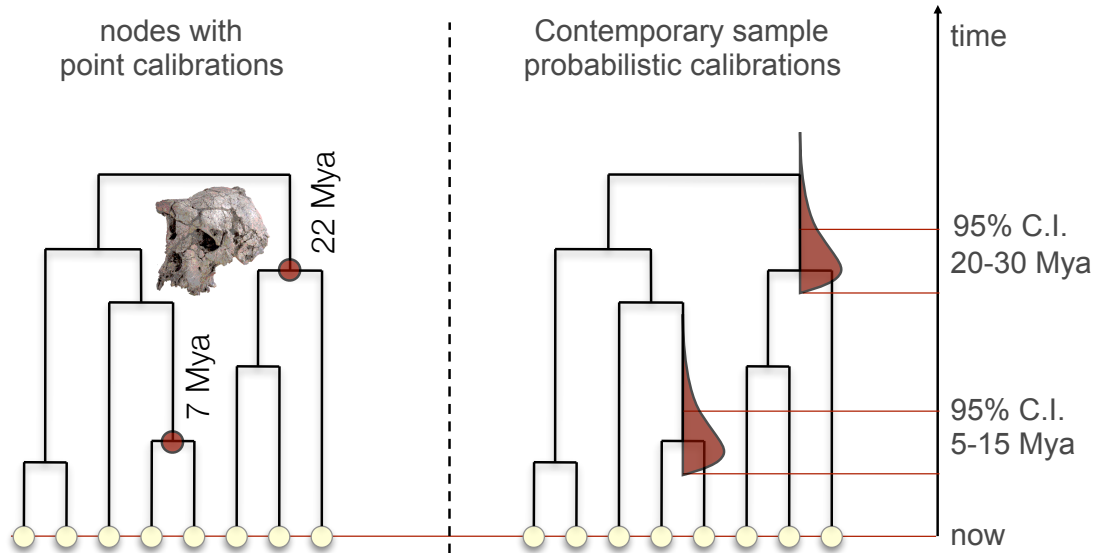
- different genes, different profiles
- variation in mutation rate?
- variation in selection
genes coding for some molecules under very strong stabilizing selection



calibrating the molecular clock



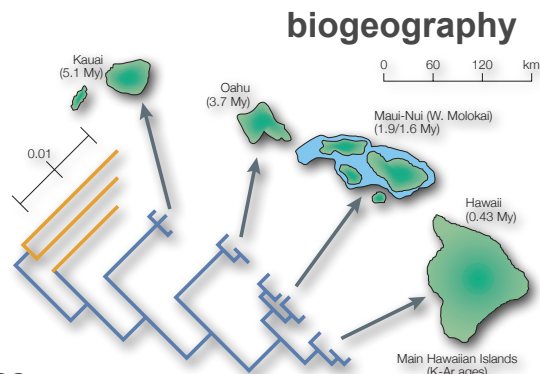
From substitution units to time units



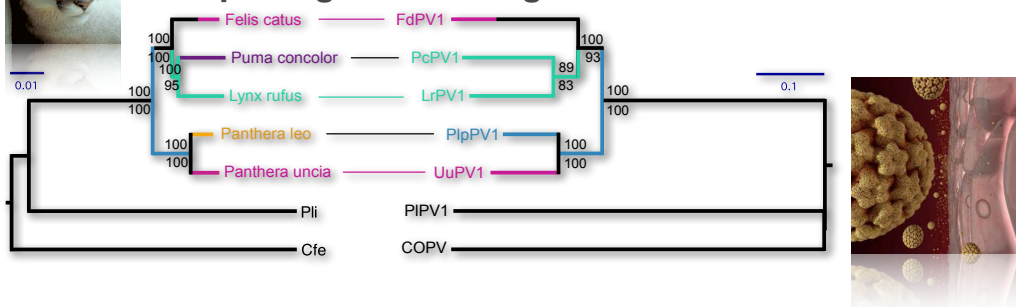
Node Calibrations



Fossils

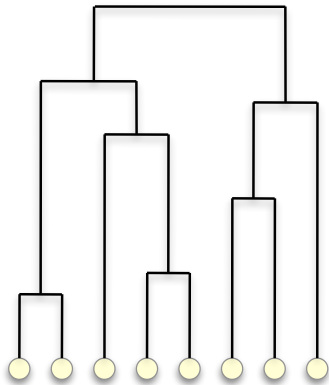


host-pathogen co-divergence

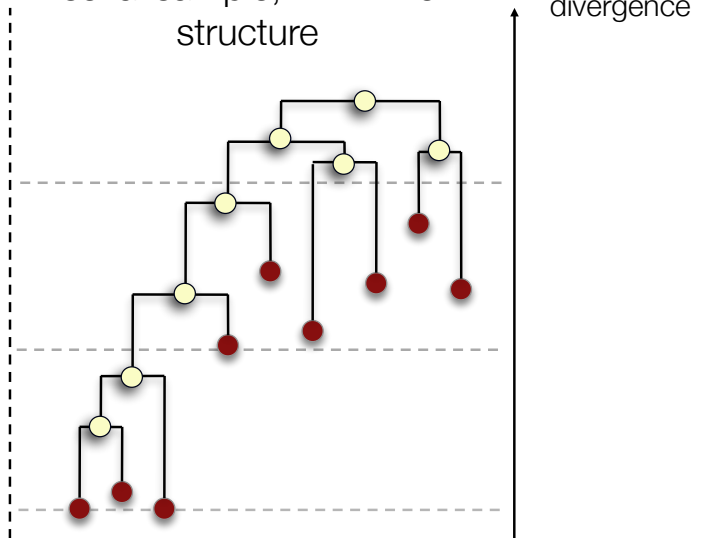


Calibration using sampling times

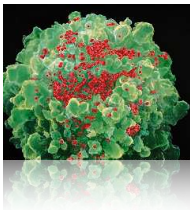
contemporary sample,
no time structure



serial sample, with time
structure



Tip calibration: two major applications



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

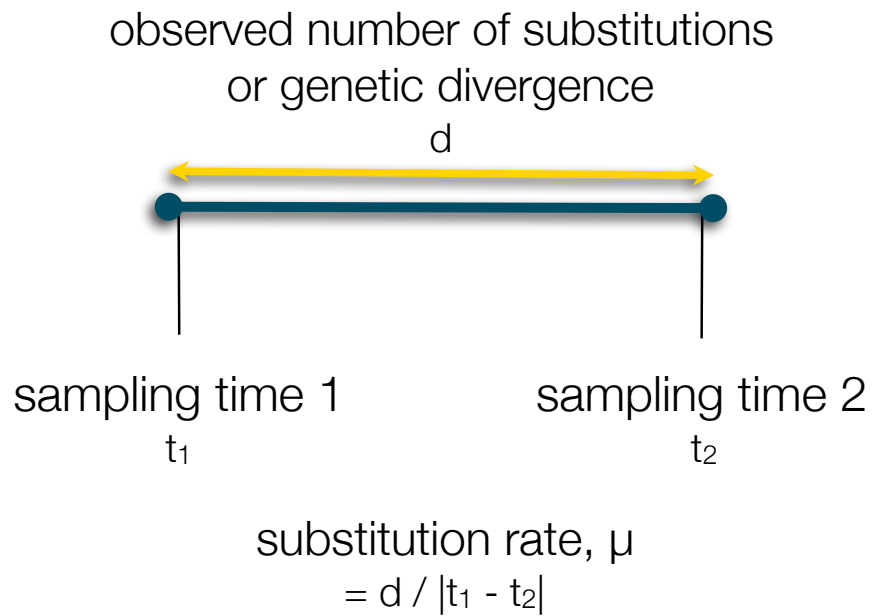
- Substitutions accumulate between the times of sampling
- Serially sampled sequences or heterochronous sequences



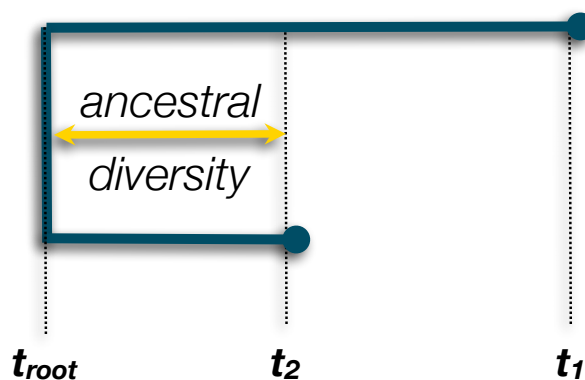
ancient DNA
data sets of
radiocarbon-dated
specimens

**Measurably evolving
population**

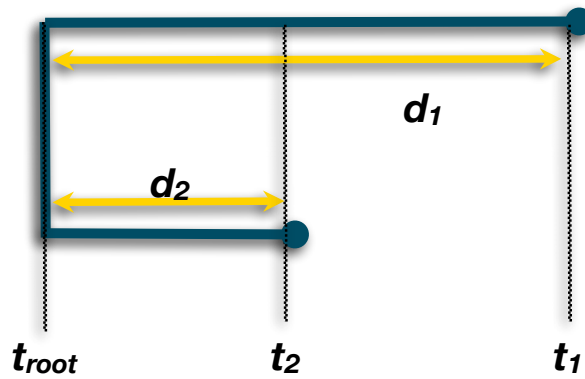
incorporating sampling time: naive method



incorporating sampling time: naive method

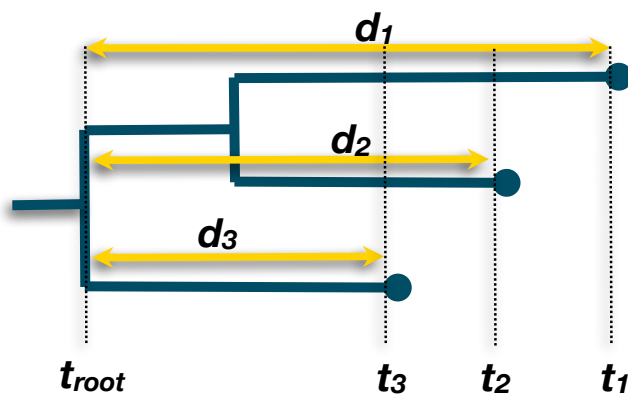


incorporating sampling time: naive method



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

linear regression



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

- can be rearranged:

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

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

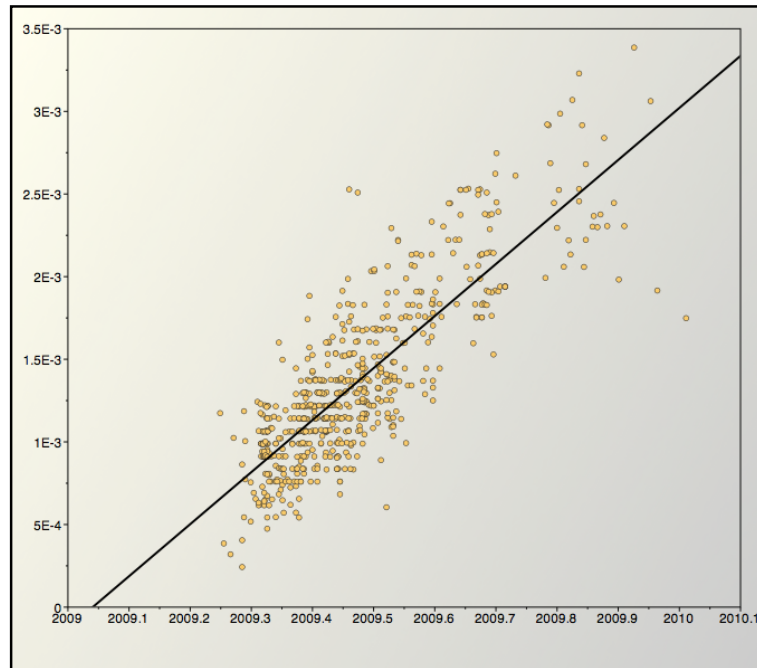
gradient is: μ

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

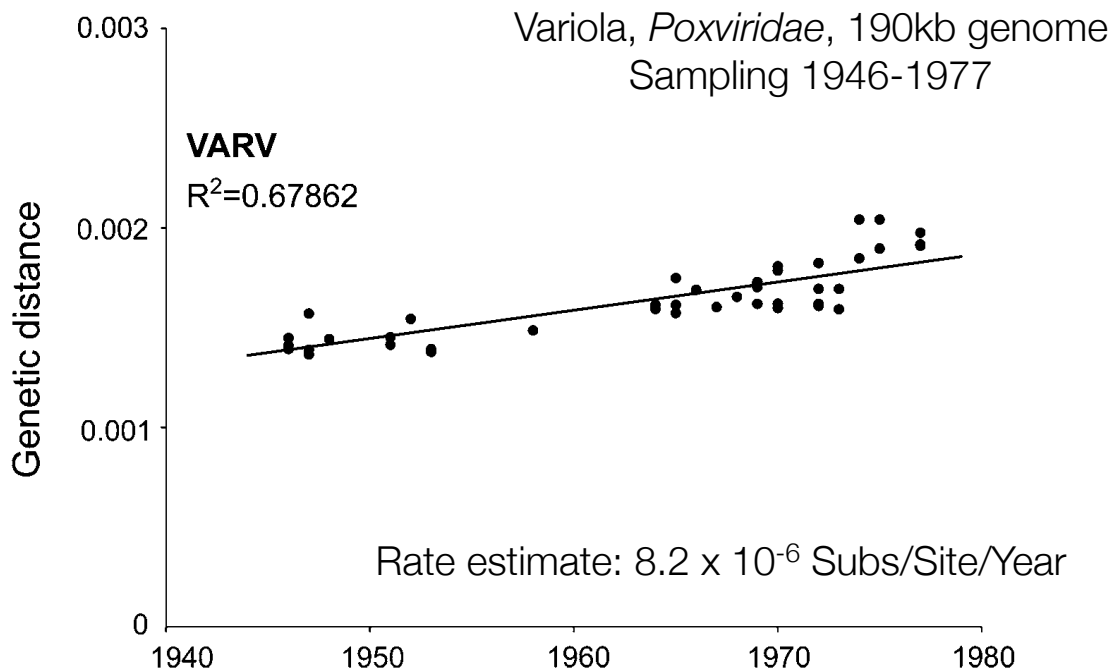
x-intercept is: t_{root}

Estimating the time-scale

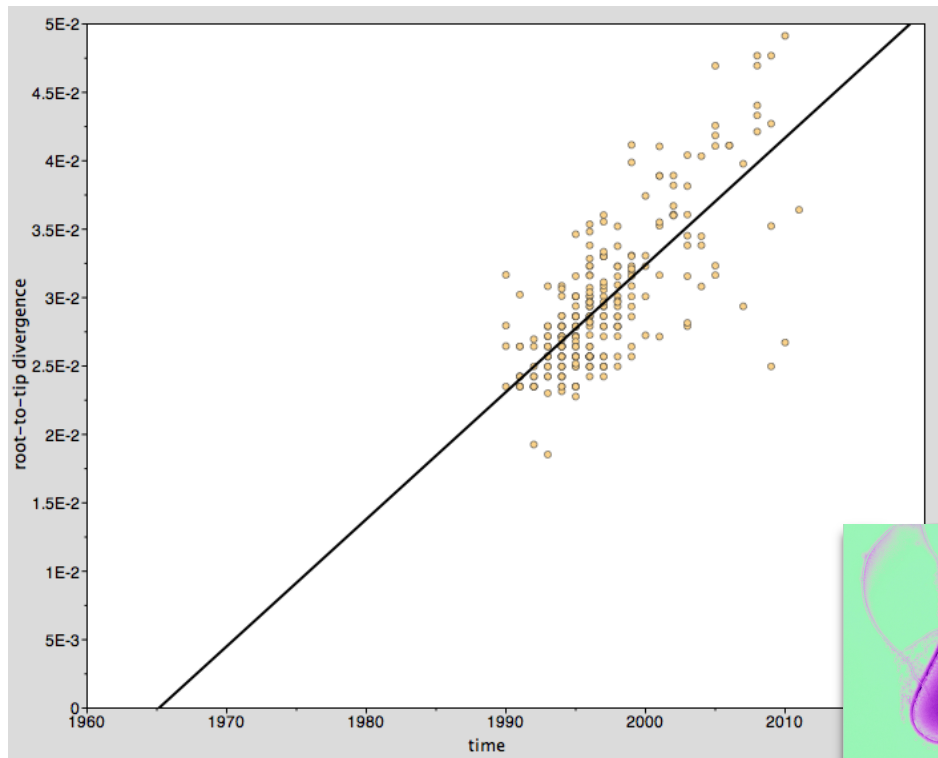
- H1N1/09 'Swine Flu'
- Rate: $3.14E^{-3}$ mutations/genomic site/year
- tMRCA: 2009.041 (15-Jan-2009)
- Correlation: 0.83
- R^2 : 0.69



A DNA virus (smallpox)



Salmonella Typhimurium

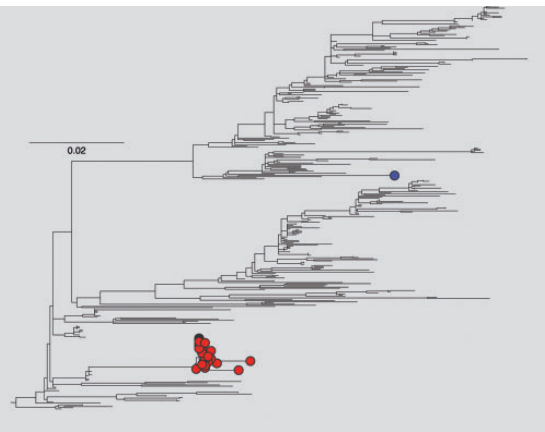
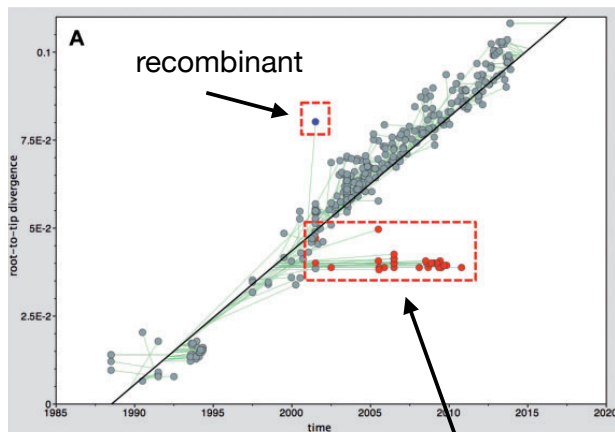


Diagnostic tool

- divergence accumulation
- outliers



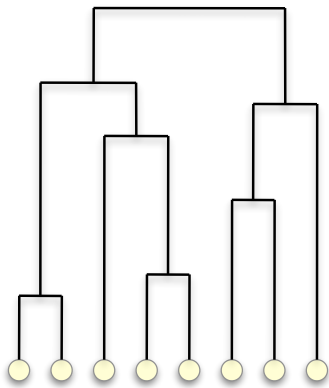
TempEst



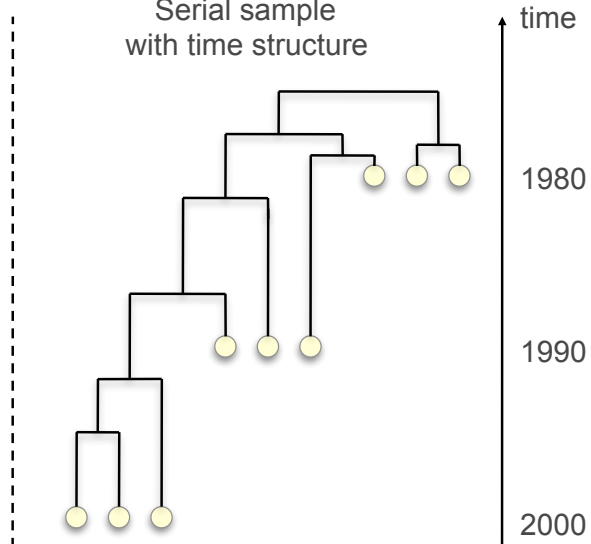
vaccine lineage

Time structure via tip calibration

Contemporary sample
no time structure

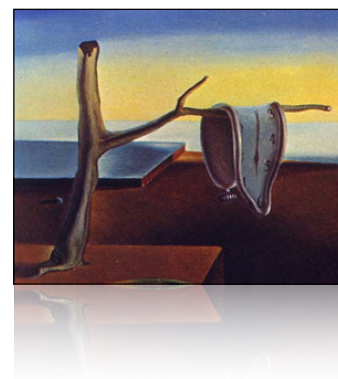


Serial sample
with time structure



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

Relaxing the molecular clock

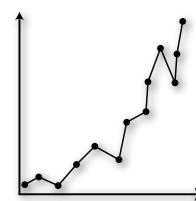
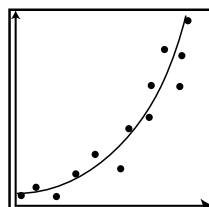
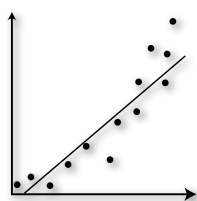


Clock versus non-clock

- 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:
Zuckermandl & 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

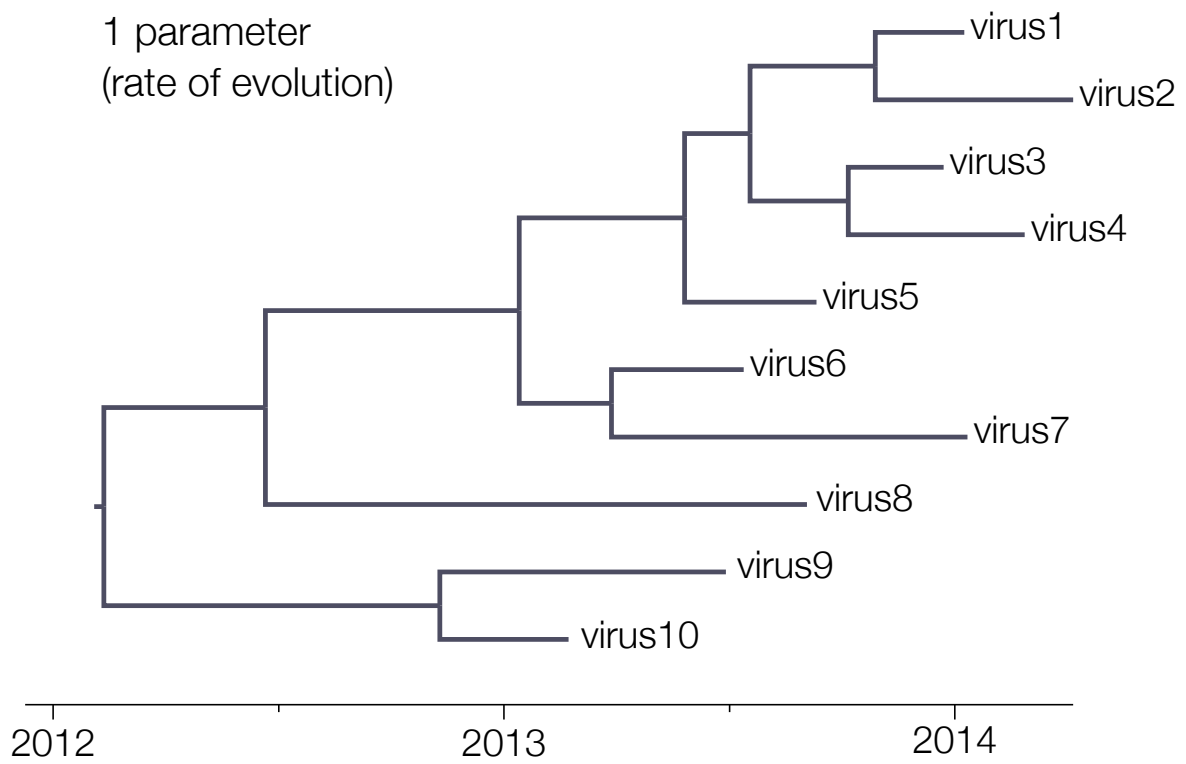
Need for a relaxed molecular clock

- 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



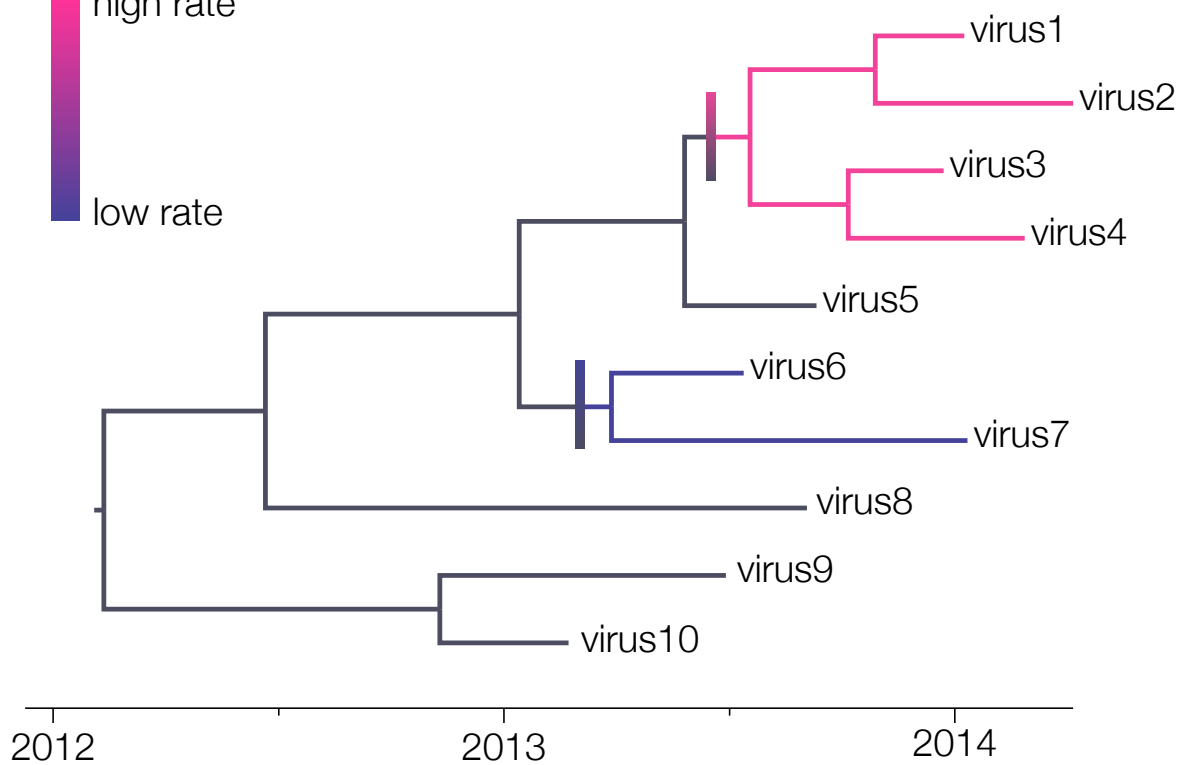
'strict' molecular clock

1 parameter
(rate of evolution)

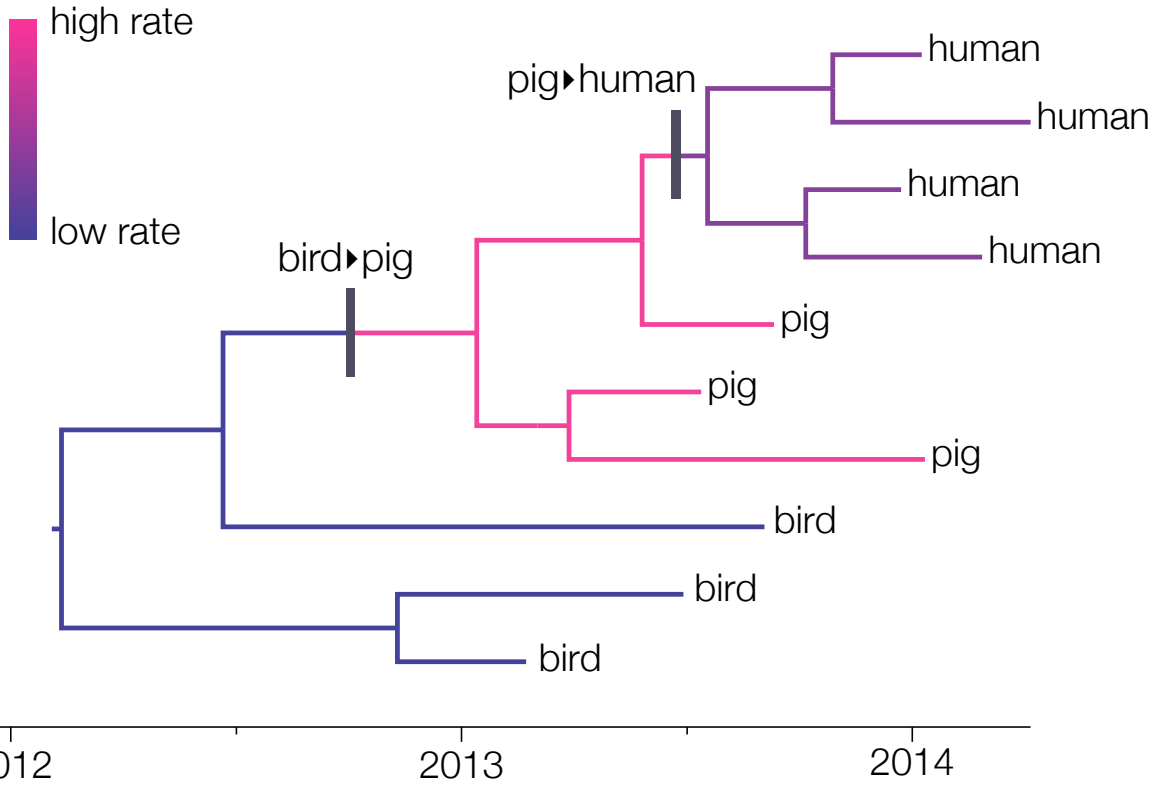


'local' molecular clock

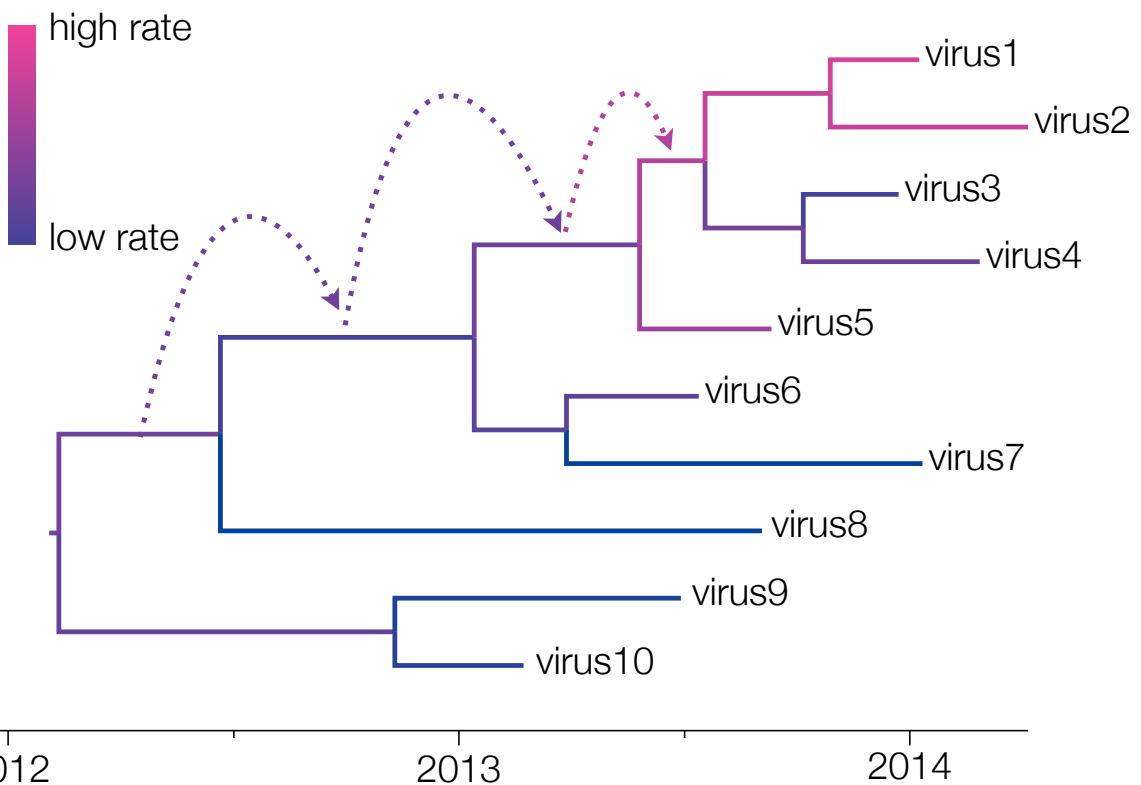
high rate
low rate

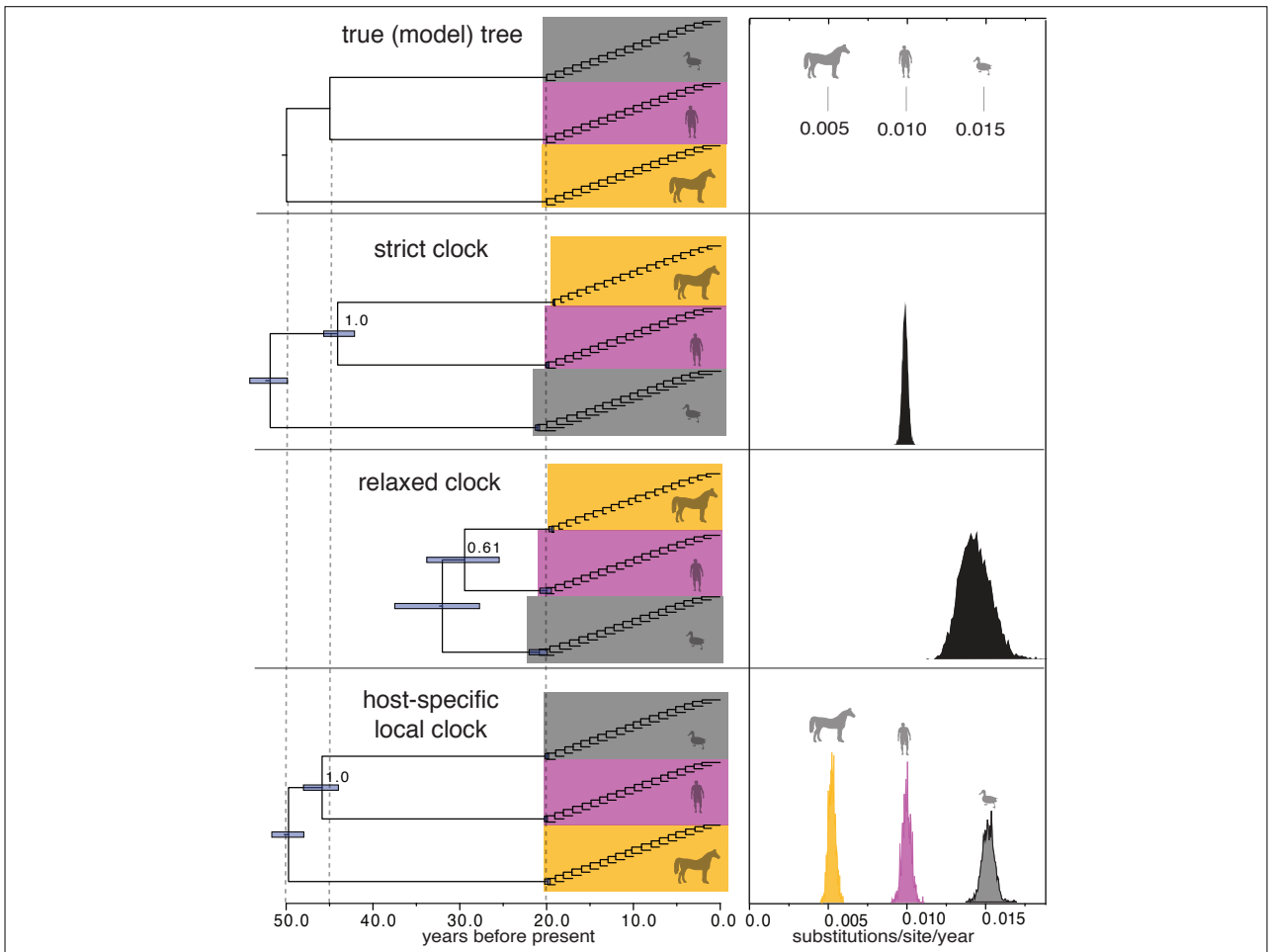


host specific local clock

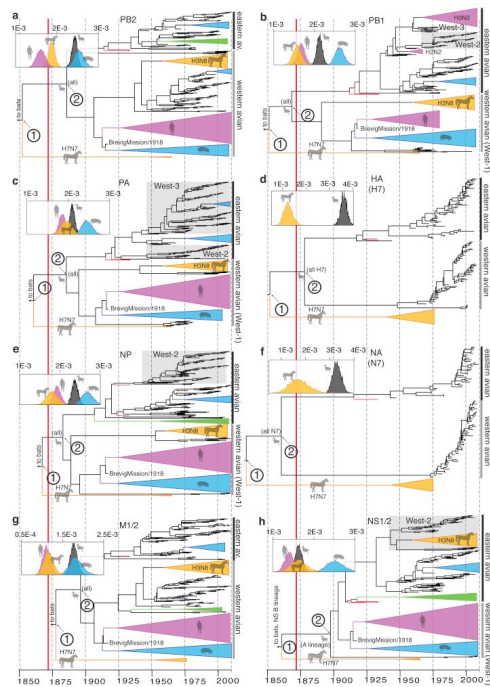
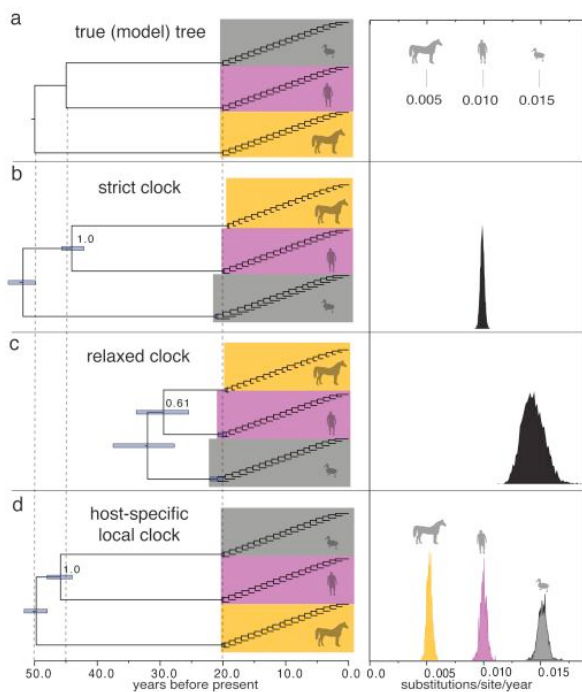


autocorrelated relaxed clock



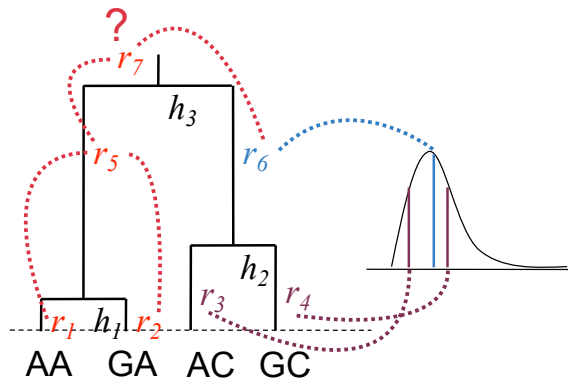


Bayesian local clocks



Autocorrelated relaxed clocks

- rates for each branch are drawn from a distribution centered 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

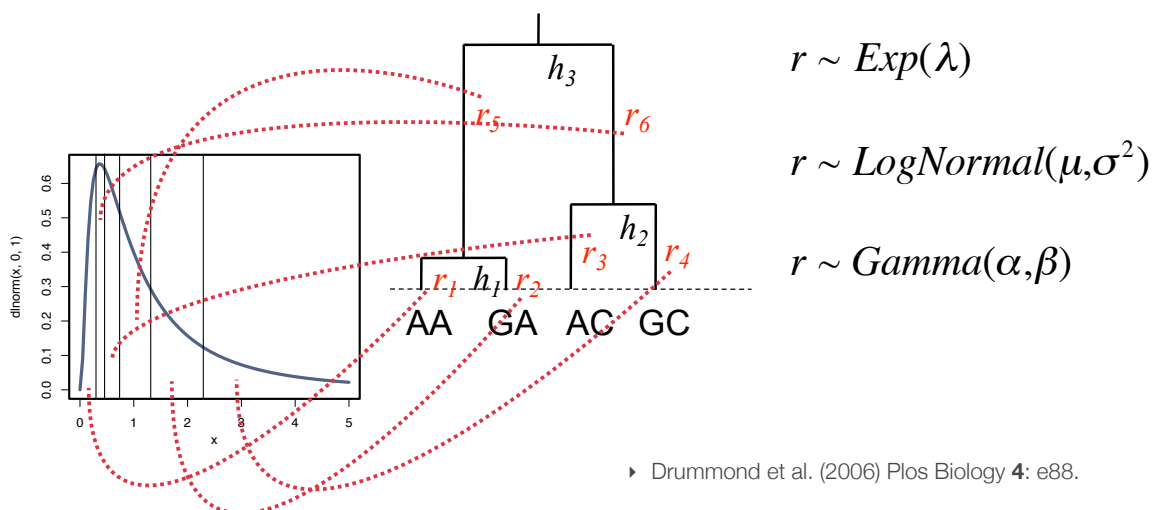


$$r_i \sim \text{LogNormal}(r_{A(i)}, \sigma^2 \Delta t_i)$$

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

Uncorrelated relaxed clocks

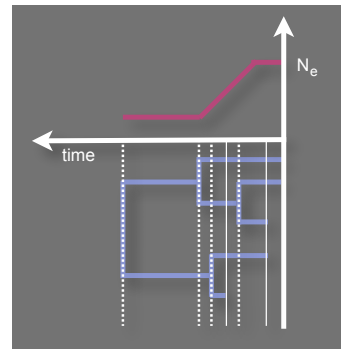
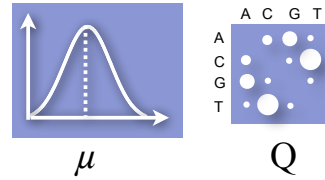
- rates for each branch are drawn independently from an identical distribution:



Bayesian evolutionary analysis sampling trees

- Given sequence data that is temporally spaced estimate true values of:

- substitution parameters (μ and Q)
- ancestral genealogy ($g = E_g, t_r$)
 - tree topology
 - dates of divergence
- population history (θ)



- Bayesian inference

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

“relaxed phylogenetics and dating with confidence”

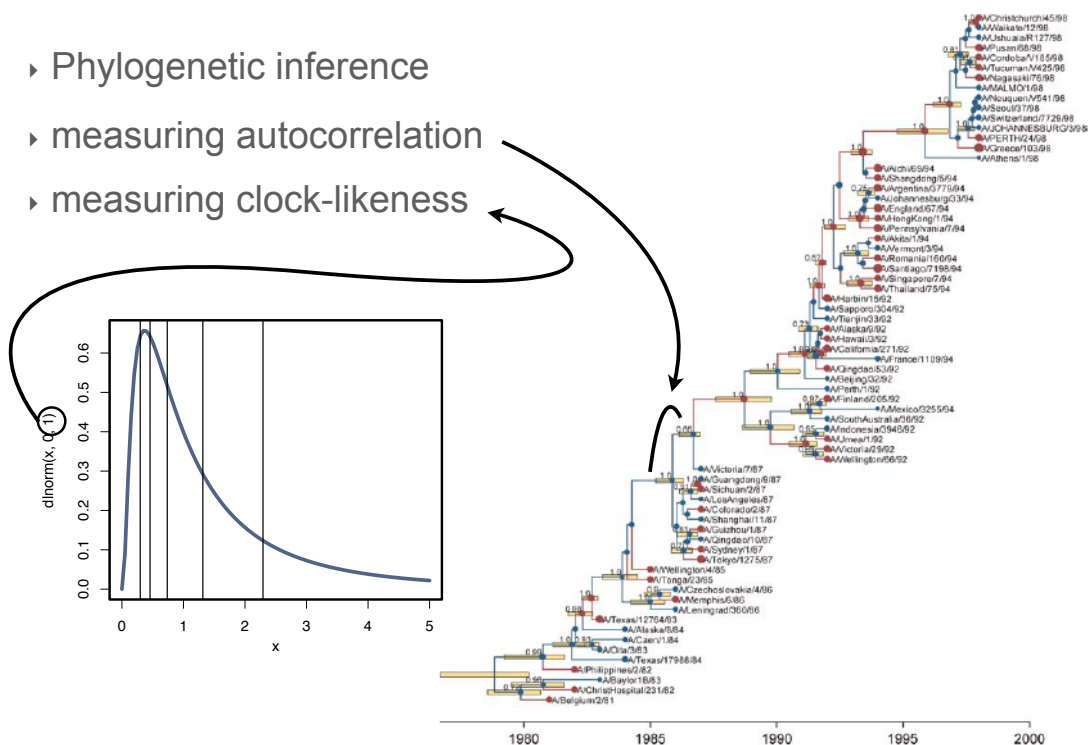
$$t = \{t_1, t_2, \dots, t_{2n-1}\}$$

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

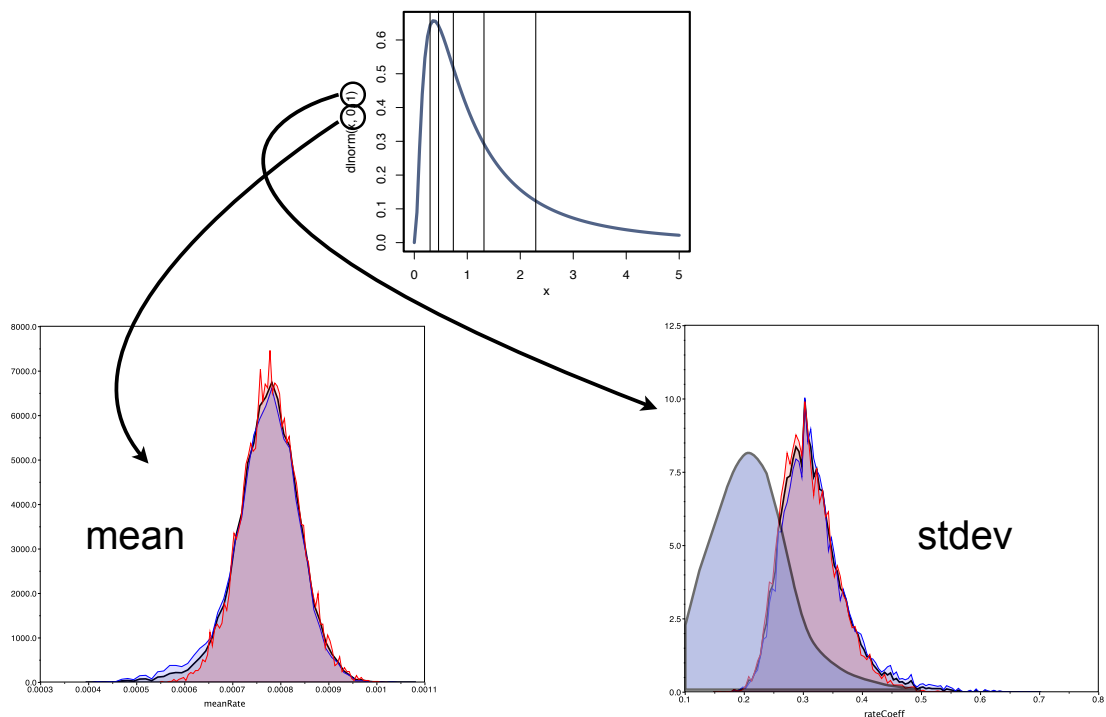
$$f(R|g) = f(R) = \prod_{i=1} \lambda e^{-\lambda t_i}$$

Uncorrelated relaxed clocks: example

- Phylogenetic inference
- measuring autocorrelation
- measuring clock-likeness



Evaluating clock-like behaviour?



Bayesian model testing

- Goal: finding the most appropriate model for your data
- Over-fitting: too many parameters, the model is too complex
- Under-fitting: too few parameters, the model is too simple
- Don't compare all possible model combinations (evolutionary model, clock models, coalescent tree prior, ...) to one another!
- Test/compare those models if that is part of the hypothesis your testing, or if your hypothesis test is sensitive to the model choice

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

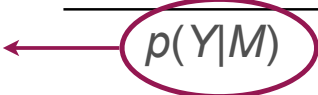
- When two models M_0 and M_1 are being compared, one defines the Bayes factor in favor of M_1 over M_0 as the ratio of their respective marginal likelihoods
- When there are unknown parameters, the Bayes Factor B_{01} has in a sense the form of a likelihood ratio

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:

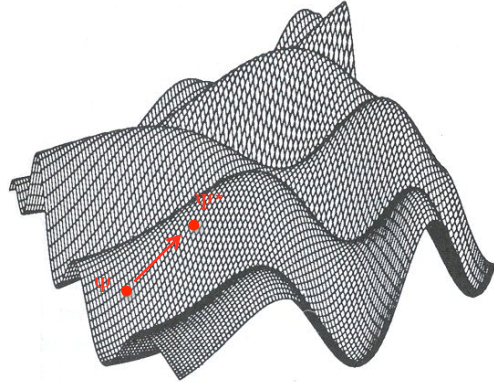
$$R = \min \left(1, \frac{p(\theta^*|D)}{p(\theta|D)} \times \frac{p(\theta|\theta^*)}{p(\theta^*|\theta)} \right)$$

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

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

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

Likelihood ratio Prior ratio Proposal ratio



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

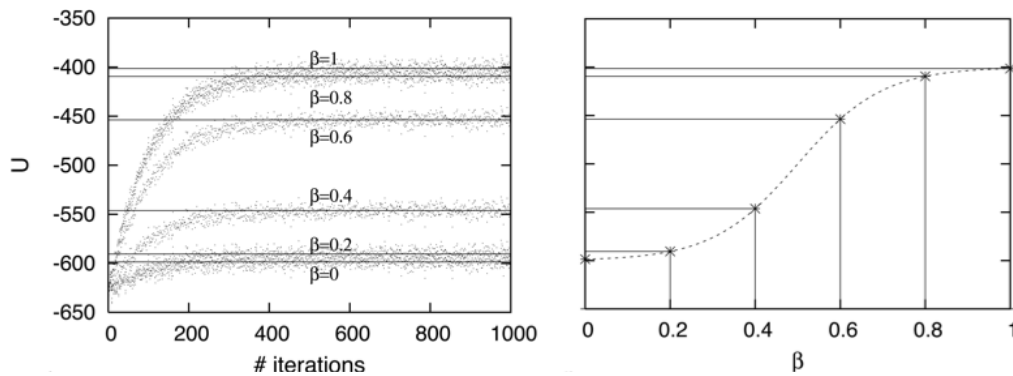
path sampling and stepping-stone sampling

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

$$q_{\beta}(\theta) = p(Y | \theta, M)^{\beta} p(\theta | M)$$

reduces to the posterior when $\beta = 1$

reduces to the prior when $\beta = 0$



path sampling and stepping-stone sampling

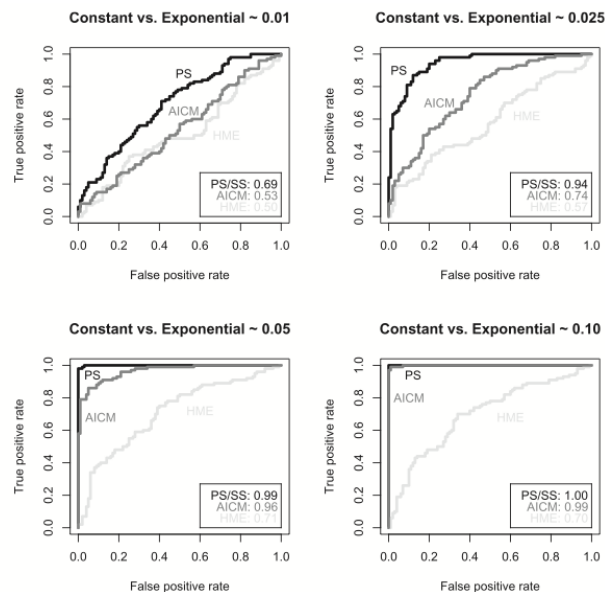


Fig. 2. Evaluation of log BF estimates using PS (SS yields an undistinguishable plot), AICM, and the HME to compare model fit, with four pairwise comparisons being shown: a constant population size versus an exponential population size with growth rates of 0.01, 0.025, 0.05, and 0.10. An increasingly 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.

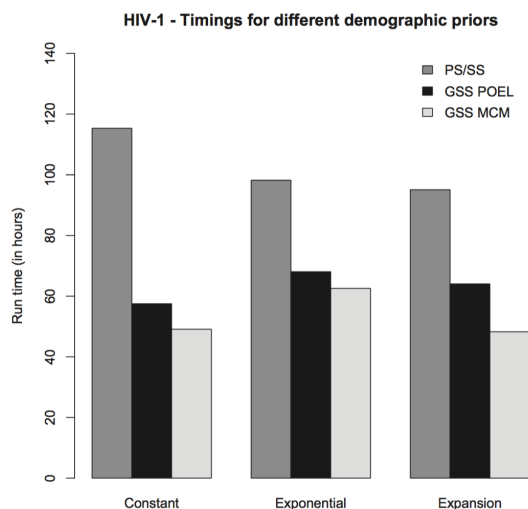
Generalised stepping-stone sampling

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

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

- 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 drastic performance increase (dependent on the actual reference/working prior)

Bayesian model selection vs model averaging

- *Test/compare those models if that is part of the hypothesis your testing, or if your hypothesis test is sensitive to the model choice*

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.

Random local clocks

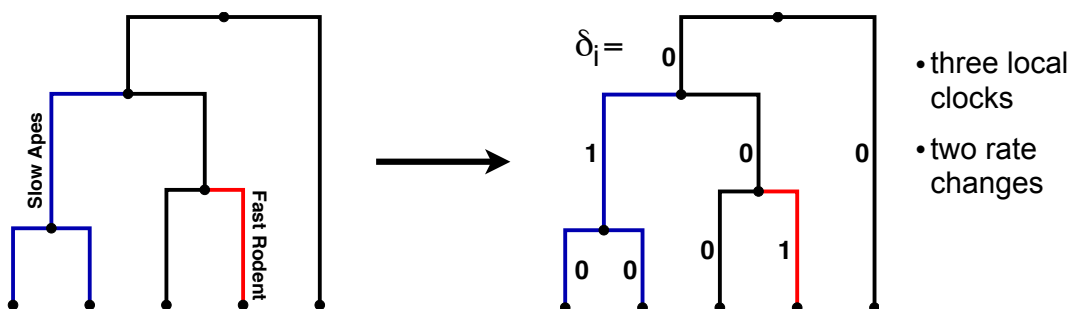
→ local clocks

- specify H_0 a priori
- problem of identifiability

→ uncorrelated relaxed clocks

- Rate changes do not necessarily occur regularly or on every branch
- Small number of significant changes

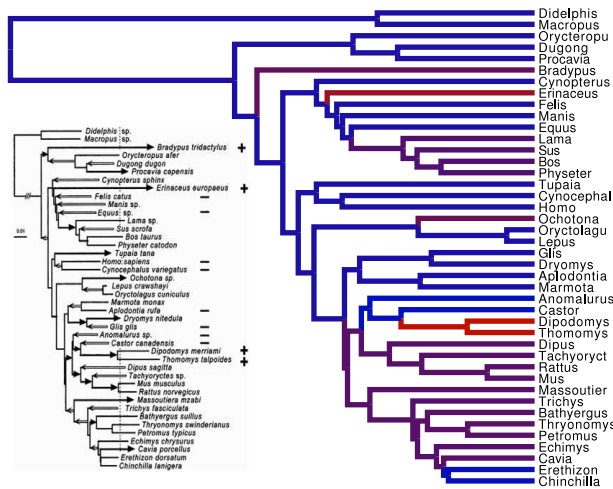
So, can we handle the uncertainty in the number and locations of a small number of local clocks?



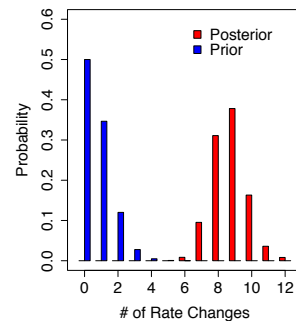
→ How to explore 2^{2n-2} clock models?

Random local clocks

- ➔ Using *Bayesian stochastic search variable selection*: formulate a prior that such that many rate changes (indicators) are 0 but allow the data to determine which ones are required to explain (most of the) rate variation using MCMC



- ➔ Three mtDNA nuclear genes from 42 mammals (Douzery, 2003)
- ➔ 5-12 local clocks



Drummond and Suchard, 2010.

Random local clocks

- ➔ Testing whether a branch accommodates a rate change using Bayes factors

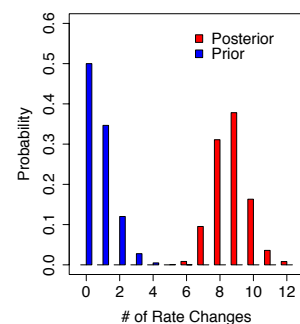
- Data D is assumed to have been arisen under one of two models, or one of two hypotheses H_1 and H_2 .

$$\text{pr}(H_k | \mathbf{D}) = \frac{\text{pr}(\mathbf{D} | H_k) \text{pr}(H_k)}{\text{pr}(\mathbf{D} | H_1) \text{pr}(H_1) + \text{pr}(\mathbf{D} | H_2) \text{pr}(H_2)}$$

so that

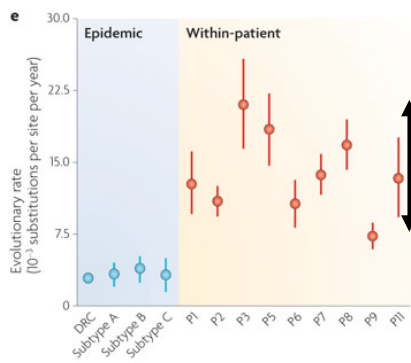
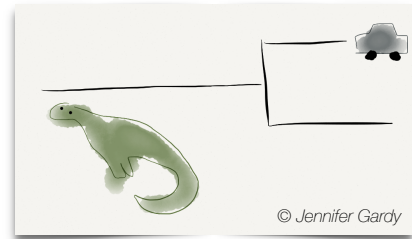
$$\frac{\text{pr}(H_1 | \mathbf{D})}{\text{pr}(H_2 | \mathbf{D})} = \frac{\text{pr}(\mathbf{D} | H_1)}{\text{pr}(\mathbf{D} | H_2)} \frac{\text{pr}(H_1)}{\text{pr}(H_2)}$$

posterior odds = Bayes factor × prior odds

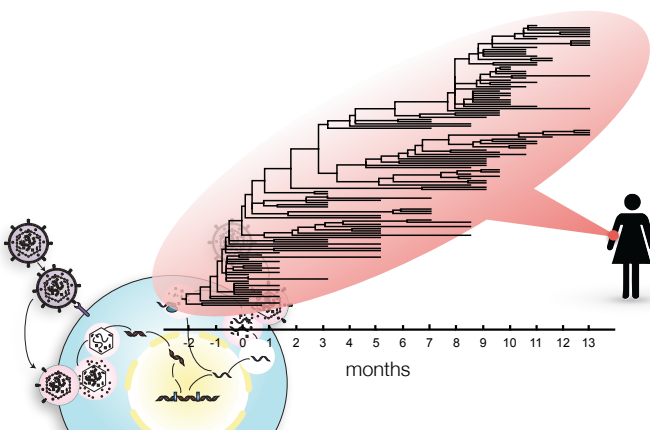


- Prior probabilities $\text{pr}(H_1)$ and $\text{pr}(H_2) = 1 - \text{pr}(H_1)$. Posterior probabilities $\text{pr}(H_1 | \mathbf{D})$ and $\text{pr}(H_2 | \mathbf{D}) = 1 - \text{pr}(H_1 | \mathbf{D})$

Extensions for testing evolutionary rate hypotheses

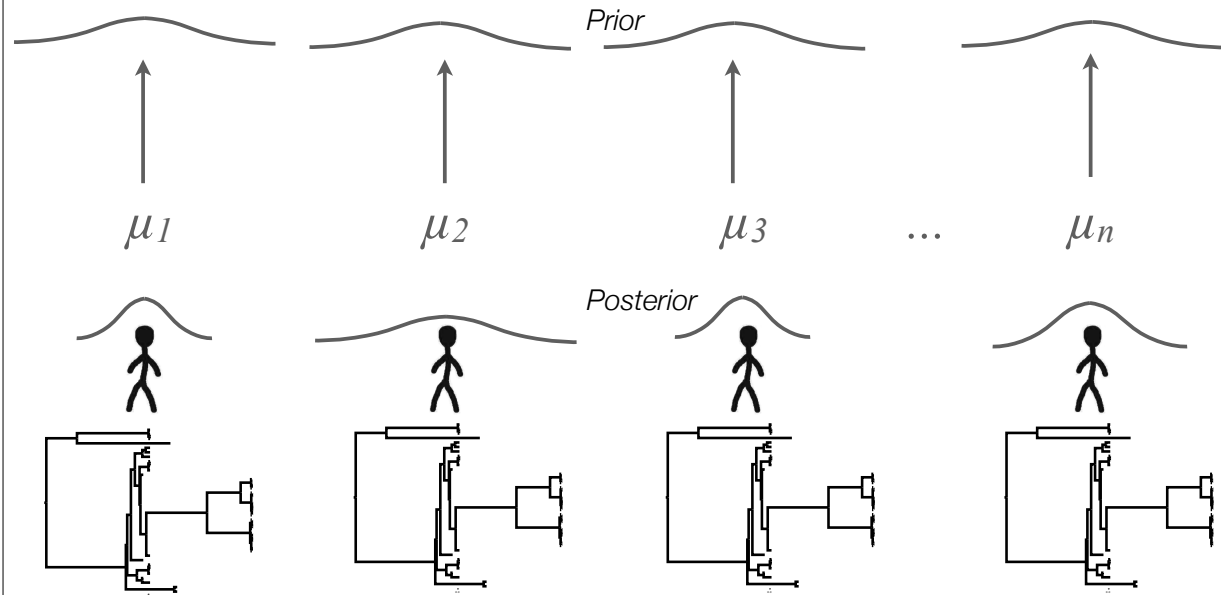


Pybus and Rambaut, NGR, 2009

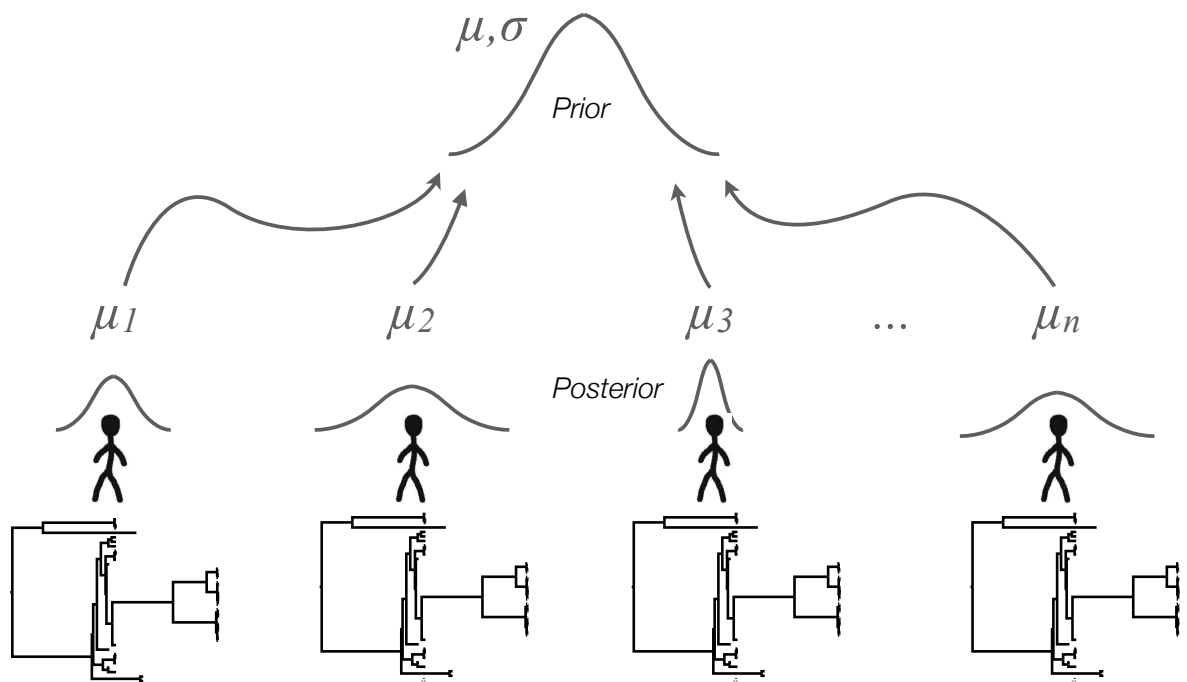


Lemey et al 2006 AIDS Rev

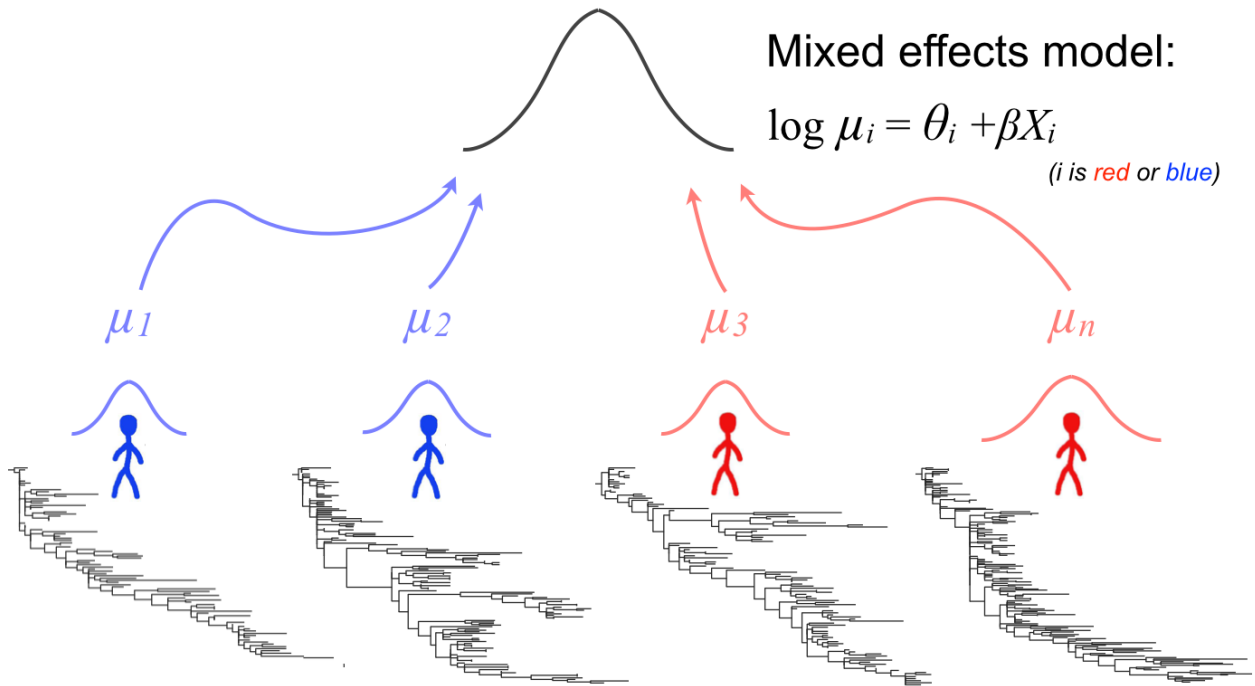
Independent parameter estimation



Hierarchical phylogenetic models

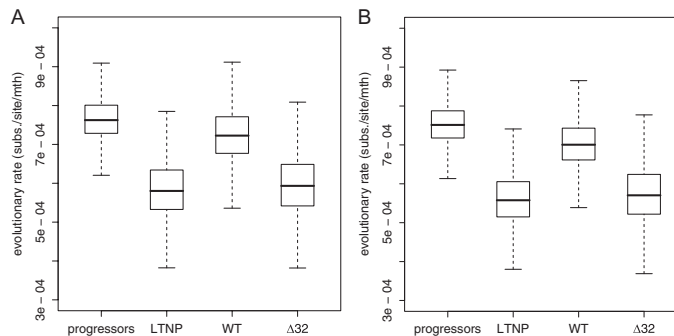


Hierarchical model with fixed effects

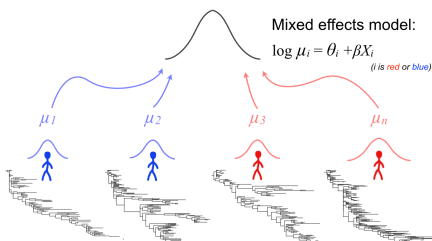


Edo-Matas et al., MBE, 2011

Hierarchical model with fixed effects

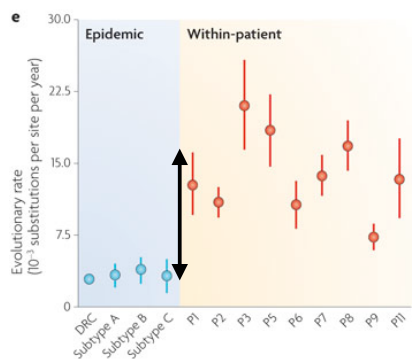


$$\log \theta_i = \beta_0 + \delta_{\text{LTNP}} \beta_{\text{LTNP}} \text{LTNP}_i + \delta_{\Delta 32} \beta_{\Delta 32} \Delta 32_i + \varepsilon_i$$

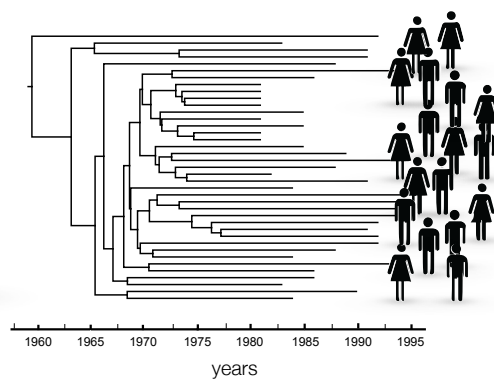
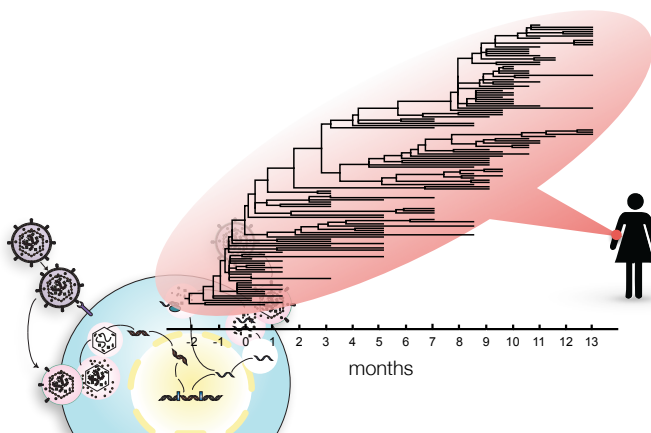


| Evolutionary Parameter | Effect Support/Size | LTNP Effect |
|------------------------------|--|-------------------------|
| Nucleotide substitution rate | Posterior probability $\delta_{\text{effect}} = 1$ | 0.72 |
| | $\text{BF}_{\text{effect}}$ | 2.6 |
| | $\beta_{\text{effect}} / \delta_{\text{effect}}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

Edo-Matas et al., MBE, 2011

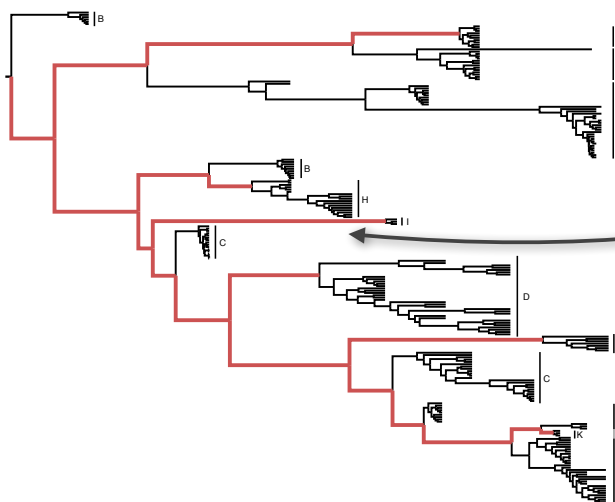


Pybus and Rambaut, NGR, 2009



Lemey et al 2006 AIDS Rev

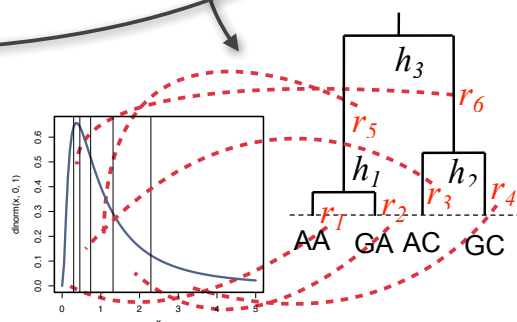
Local clocks with random effects



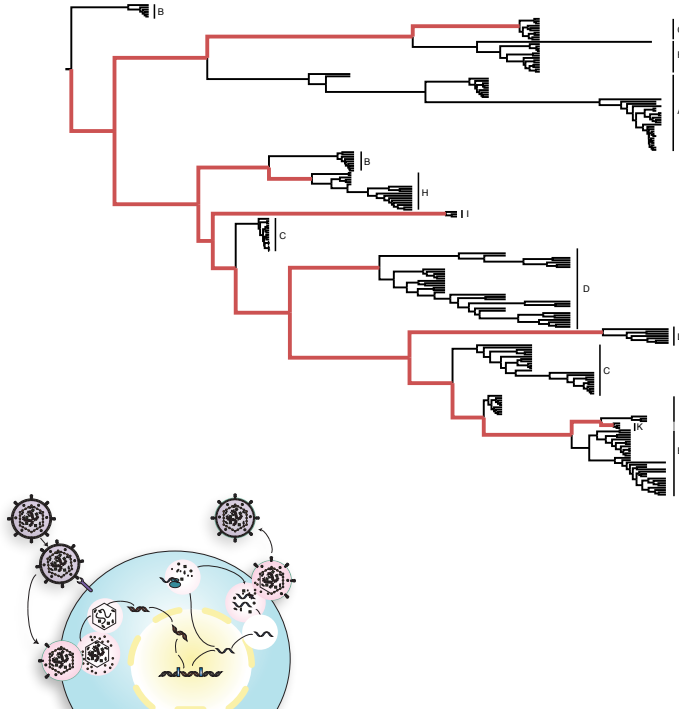
Mixed effects model:

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

(i is red or black)



Rates of HIV evolution within and between hosts

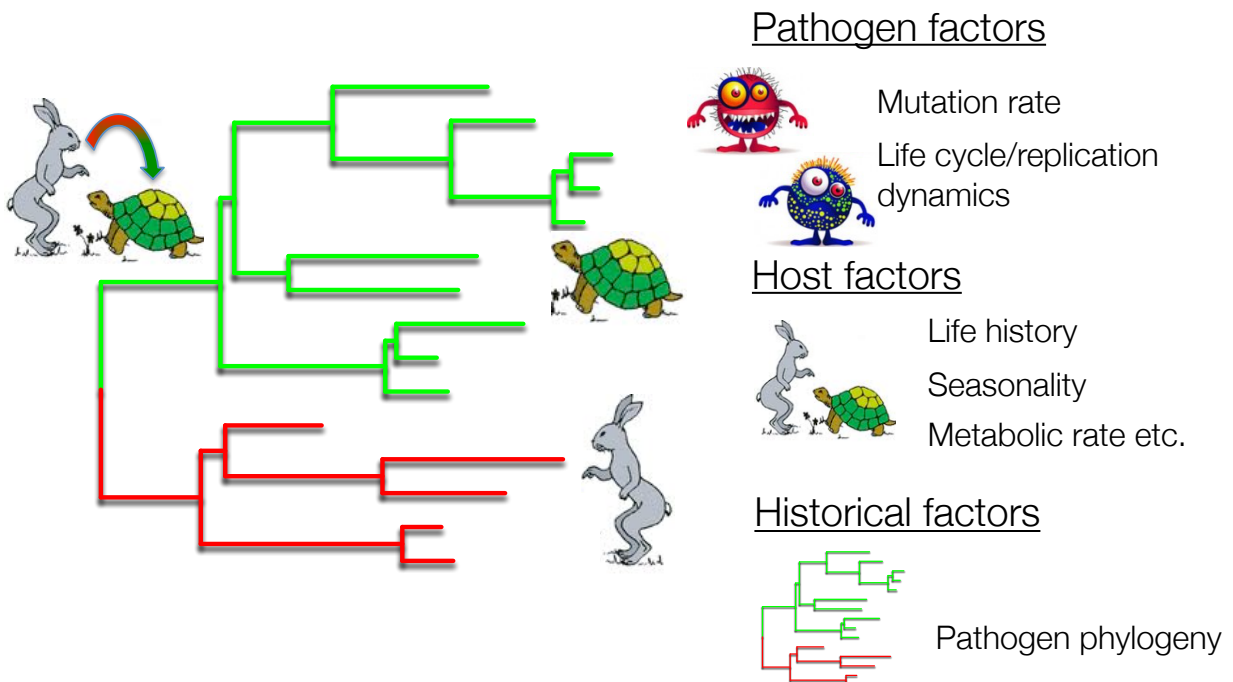


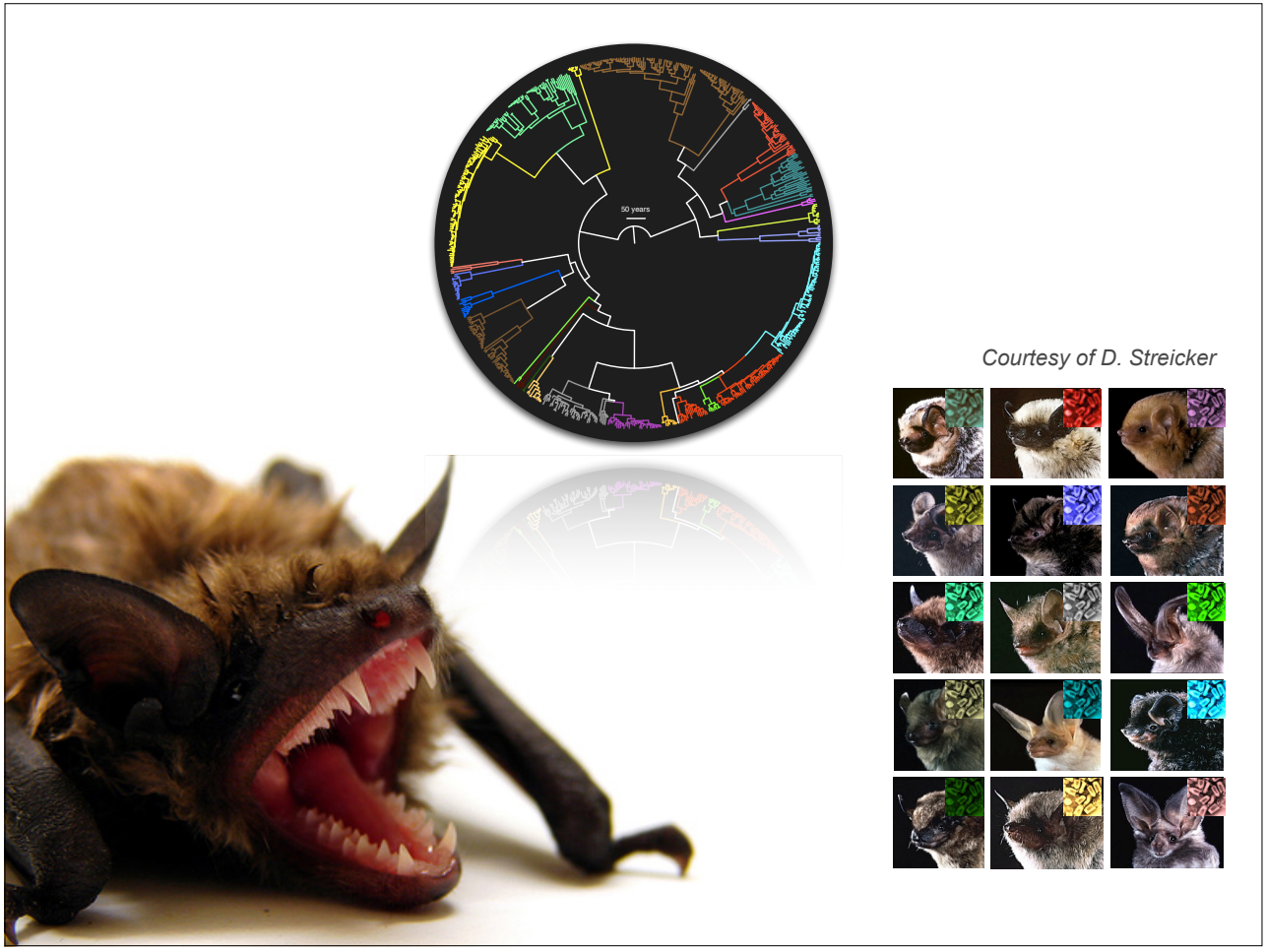
Mixed effects model:
 $\log \mu_i = \theta_i + \beta X_i$

| | <i>pol</i> | <i>env</i> |
|--|---------------------|-----------------------|
| Rate (10^{-3} subst./site/yr) | | |
| $X_i=0$ (within host) | 5.70 (4.02-6.21) | 10.37 (8.06-12.76) |
| $X_i=1$ (transmitted lineage) | 2.21 (1.57-2.99) | 3.80 (2.32-5.20) |
| ln Bayes factor ($rate_{transmitted} < rate_{within}$) | | |
| | >7.50 | >6.29 |

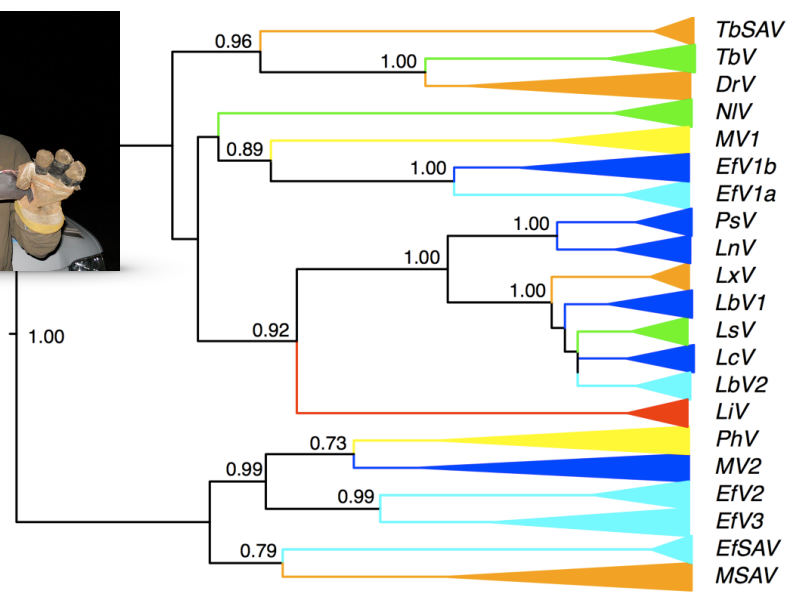
Vrancken et al., PLoS Comp Bio, 2014

What drives the tempo of pathogen evolution?



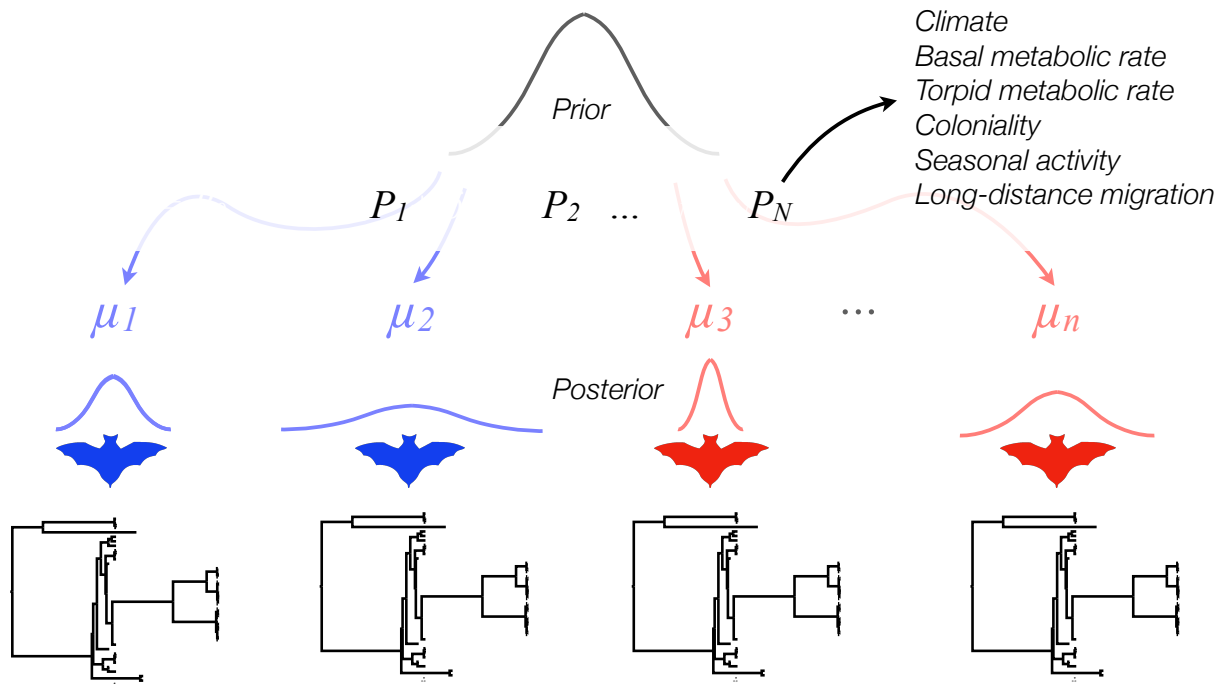


Bat rabies virus evolutionary rates



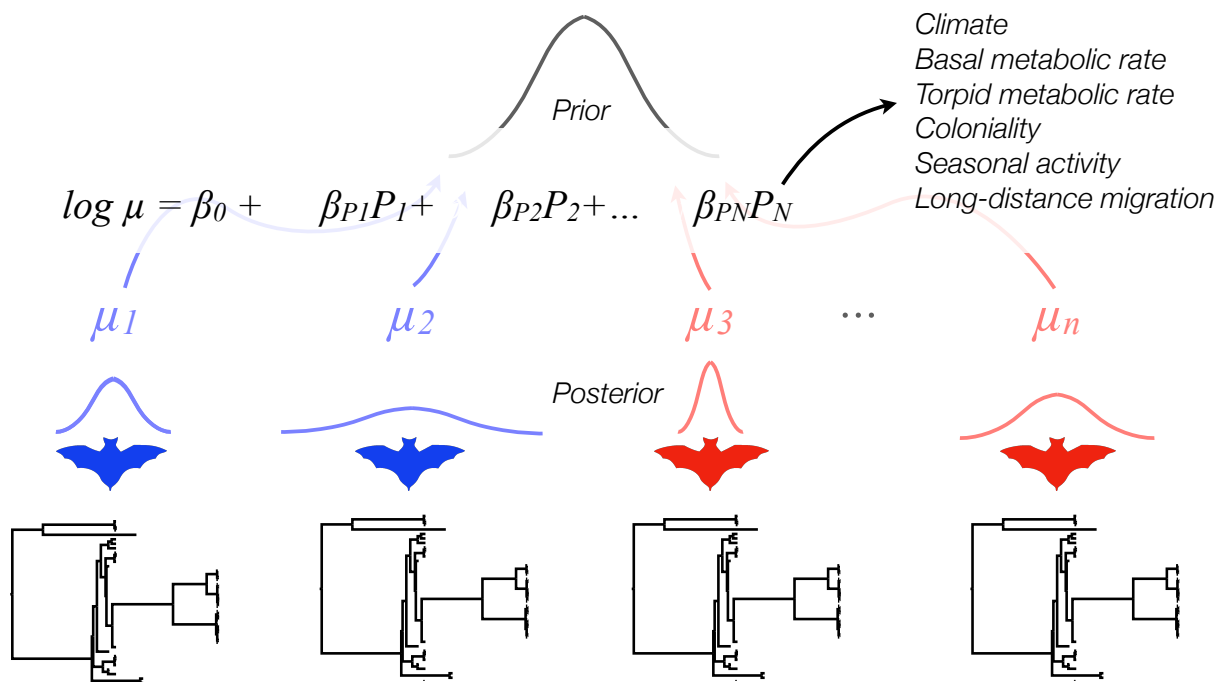
- 8.79e-5 - 4.22e-4
- 4.23e-4 - 7.55e-4
- 7.56e-4 - 1.09e-3
- 1.10e-3 - 1.42e-3
- 1.43e-3 - 1.76e-3
- 1.77e-3 - 2.09e-3

Fixed-effect hierarchical phylogenetic models



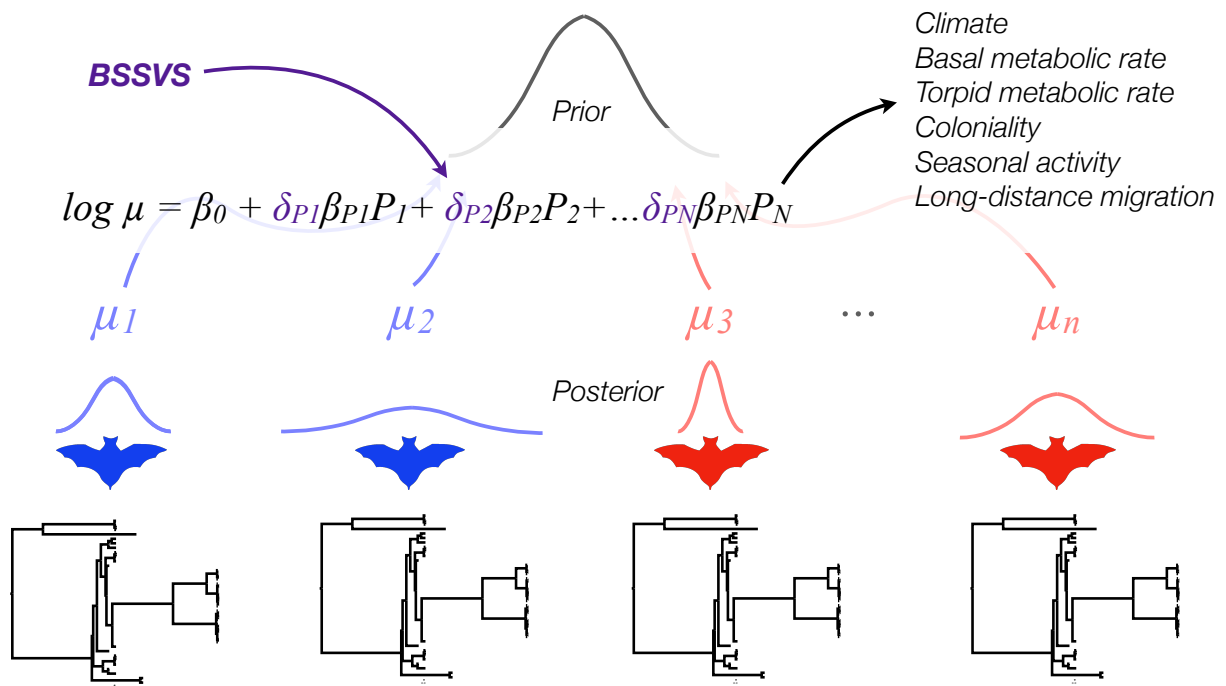
Edo-matas et al., 2011. *MBE*

Fixed-effect hierarchical phylogenetic models



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

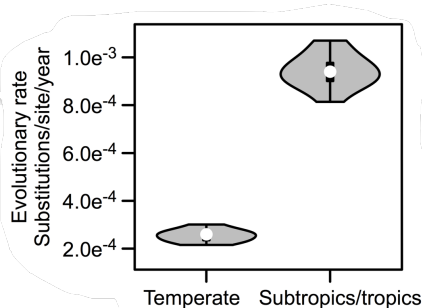
Fixed-effect hierarchical phylogenetic models



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

Bat rabies virus evolutionary rates

| Predictor | Bayes factor | β (95% HPD) $\delta = 1$ |
|-------------------------|--------------|----------------------------------|
| Climate | 466.54 | ~1.2 |
| Basal metabolic rate | 0.82 | ~0.0 |
| Torpid metabolic rate | 1.00 | ~0.0 |
| Coloniality | 0.46 | ~-0.2 |
| Seasonal activity | 0.46 | ~0.2 |
| Long-distance migration | 0.69 | ~-0.4 |



Streicker et al., 2012. *PLoS Pathogens*