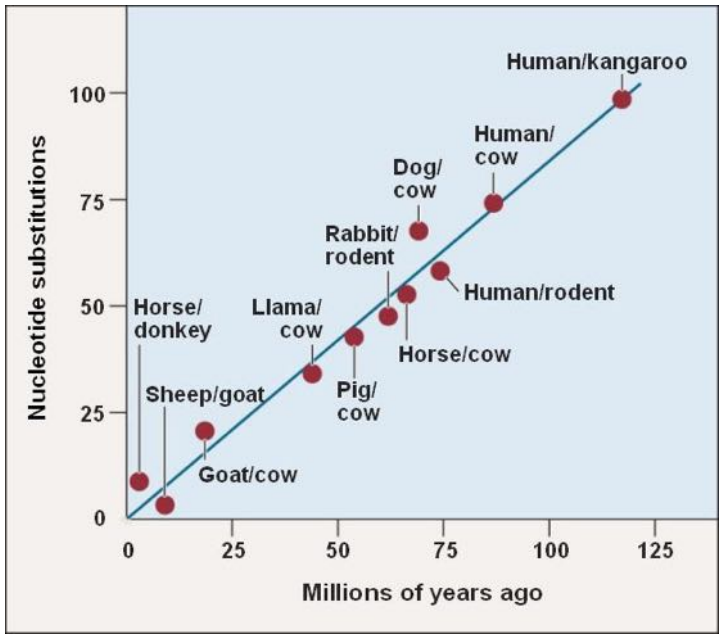
An introduction to molecular clocks in pathogen evolution

Roman Biek
(Roman.Biek@glasgow.ac.uk)

Institute of Biodiversity, Animal Health & Comparative Medicine

Molecular clocks

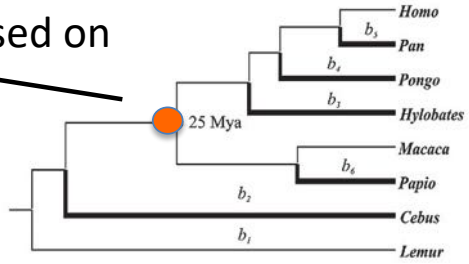


- Originally described for vertebrates (Zuckerkandl &Pauling 1962):

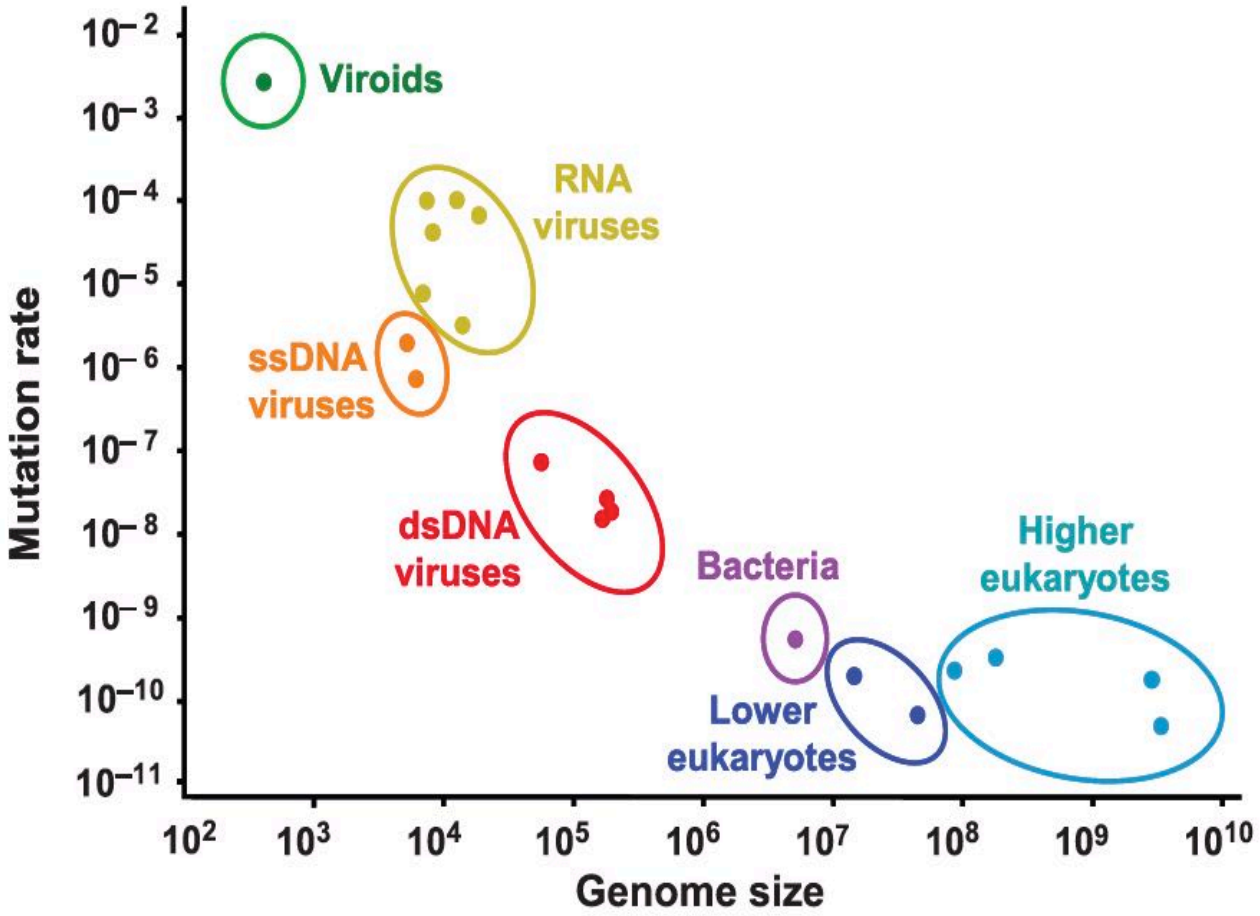
  "Genetic divergence between species increases linearly since the time these species shared a common ancestor"

- Originally relied on dates from the fossil record to calibrate tree nodes

Age information based on fossil records

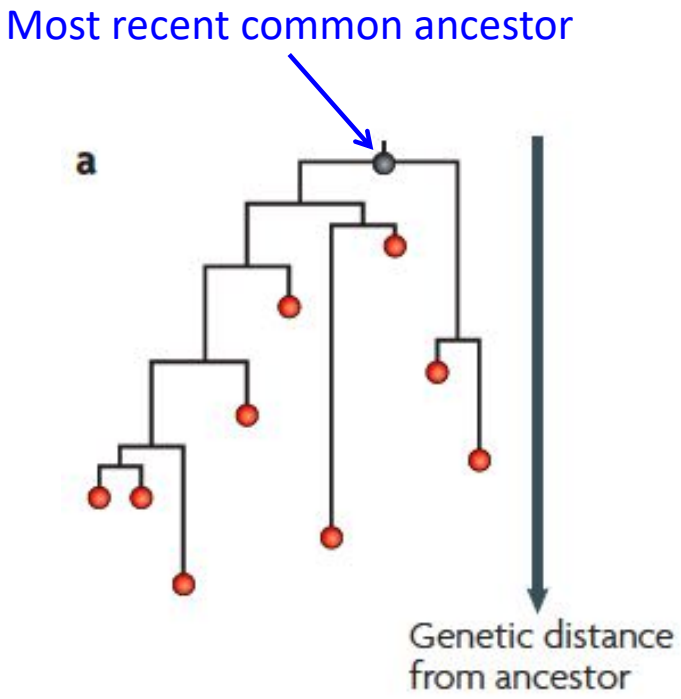


High mutation rate in microbial populations results in observable evolution



Gago et al 2009, Science

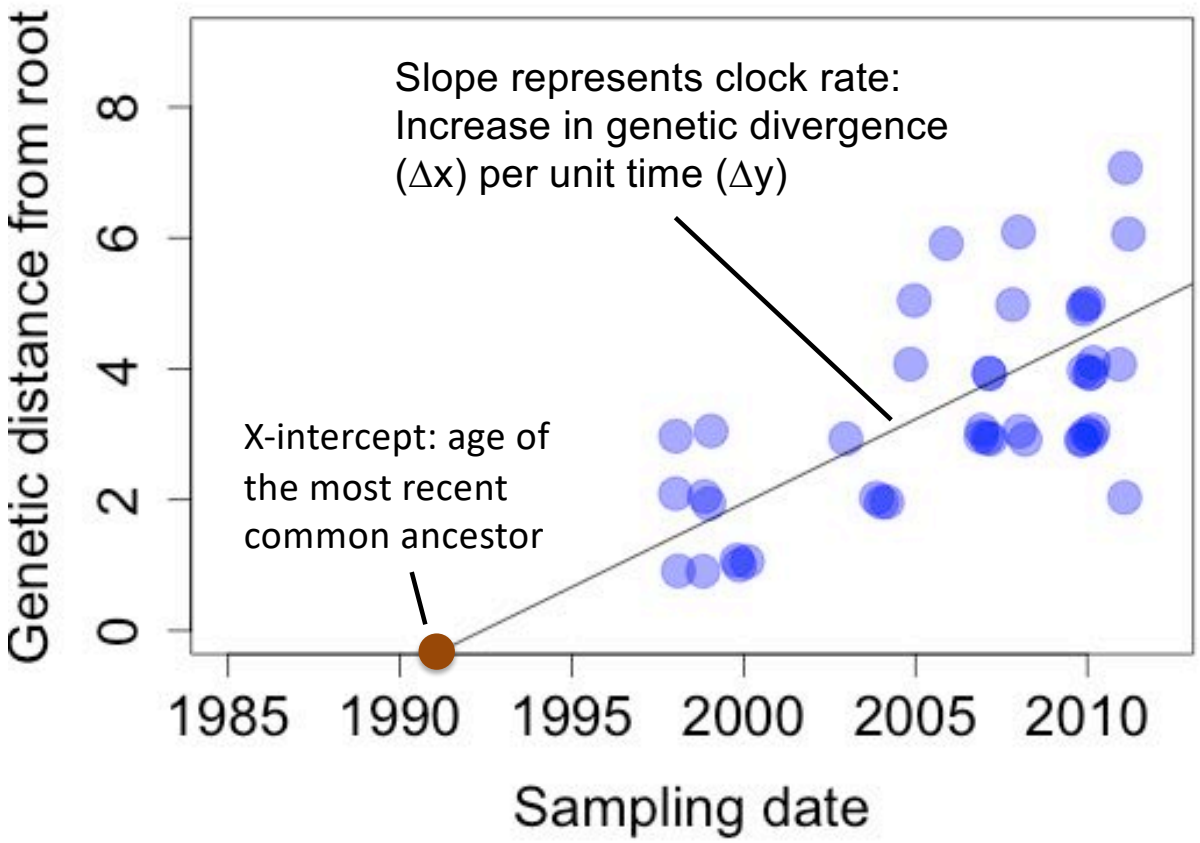
Molecular clocks



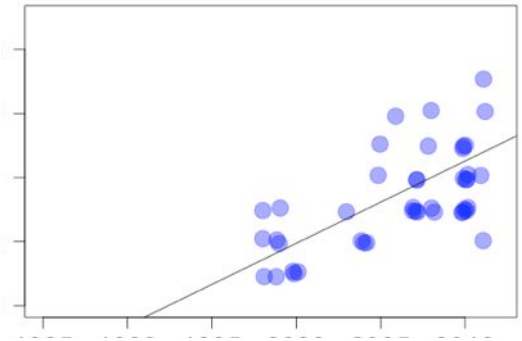
Pybus & Rambaut 2008, Nat Rev Genetics

Are genetic and temporal distances correlated?

Molecular clocks for pathogens



Molecular clocks for pathogens

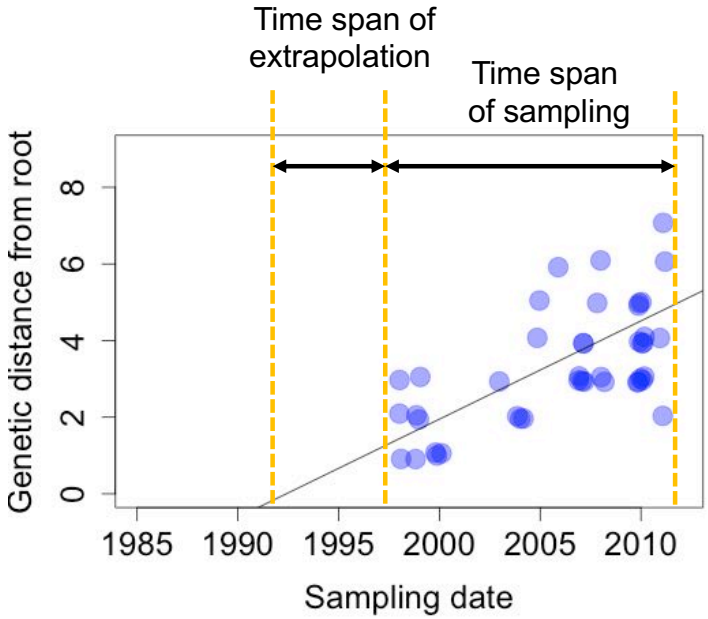


If positive relationship between genetic divergence and sampling date is found, population is said to be **measurably evolving** (Drummond et al. 2003)

Increased divergence likely caused by **mutations that are effectively neutral**

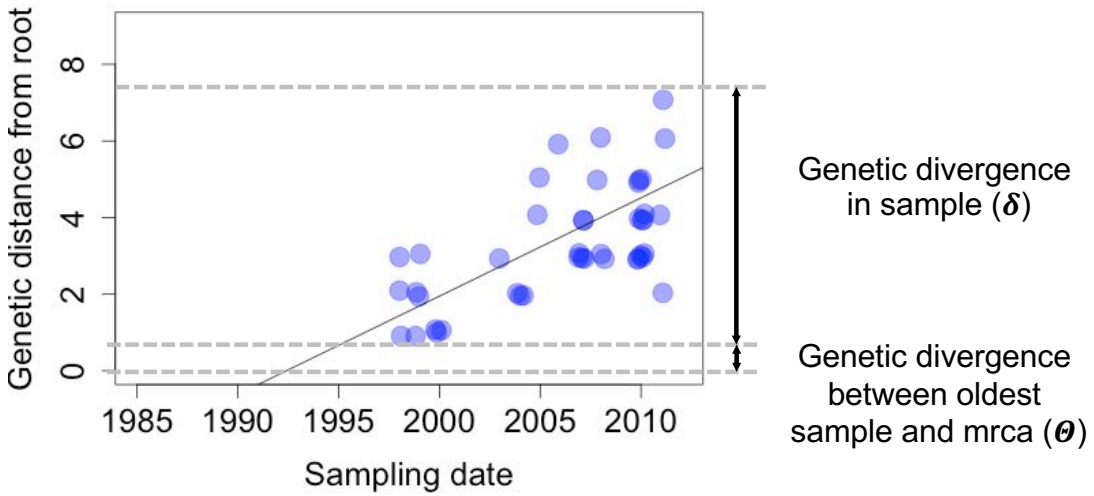
=> expect limited fitness consequences (at least in the short term)

Testing for clock-like signal using root-to-tip regression



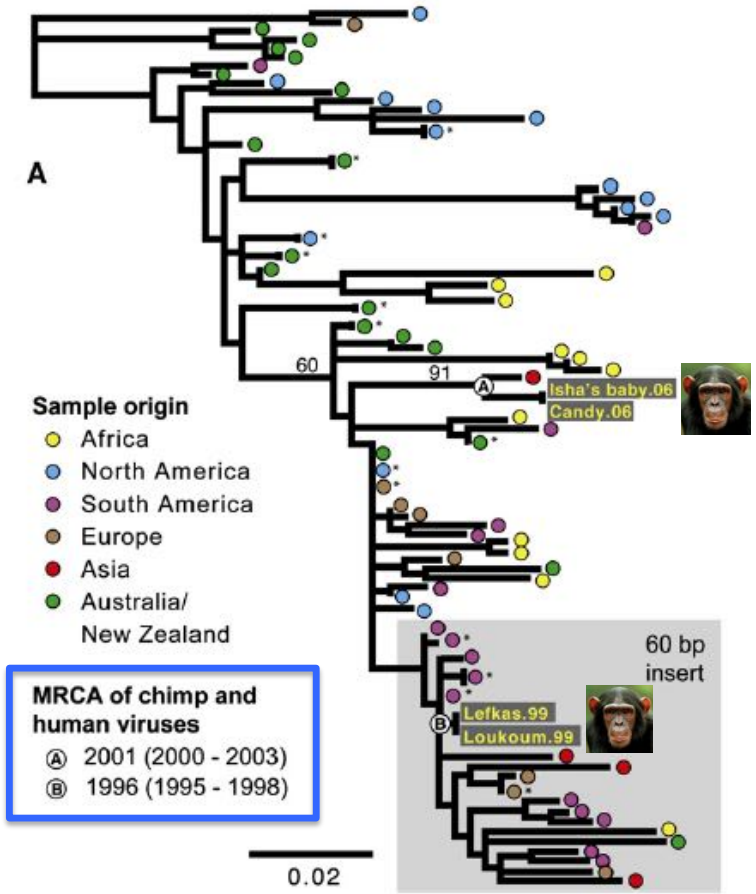
Ideally, time span of sampling > time span of extrapolation

Testing for clock-like signal using root-to-tip regression



Ideally, genetic divergence in sample > genetic divergence between oldest sample and mrca

Dating internal nodes within a phylogeny

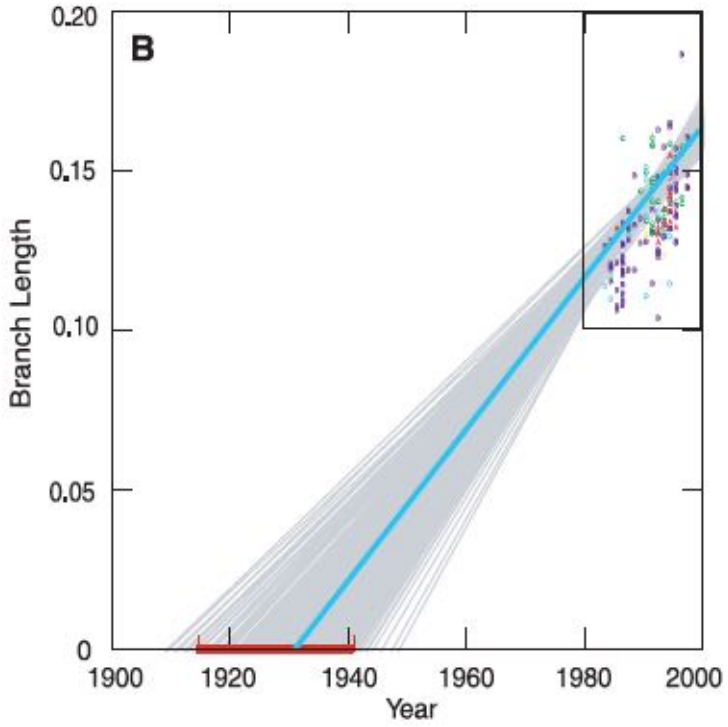


Transmission of a human respiratory virus (HRSV) from humans to chimps: How long ago?

Köndgen 2008, Current Biology

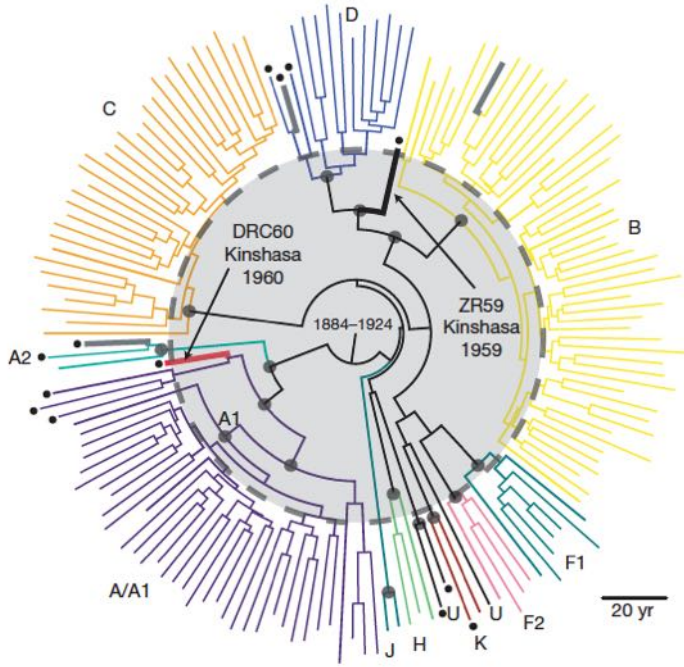
Dating the origin of the HIV pandemic

Early estimates: 1930's



Korber et al. 2000, Science

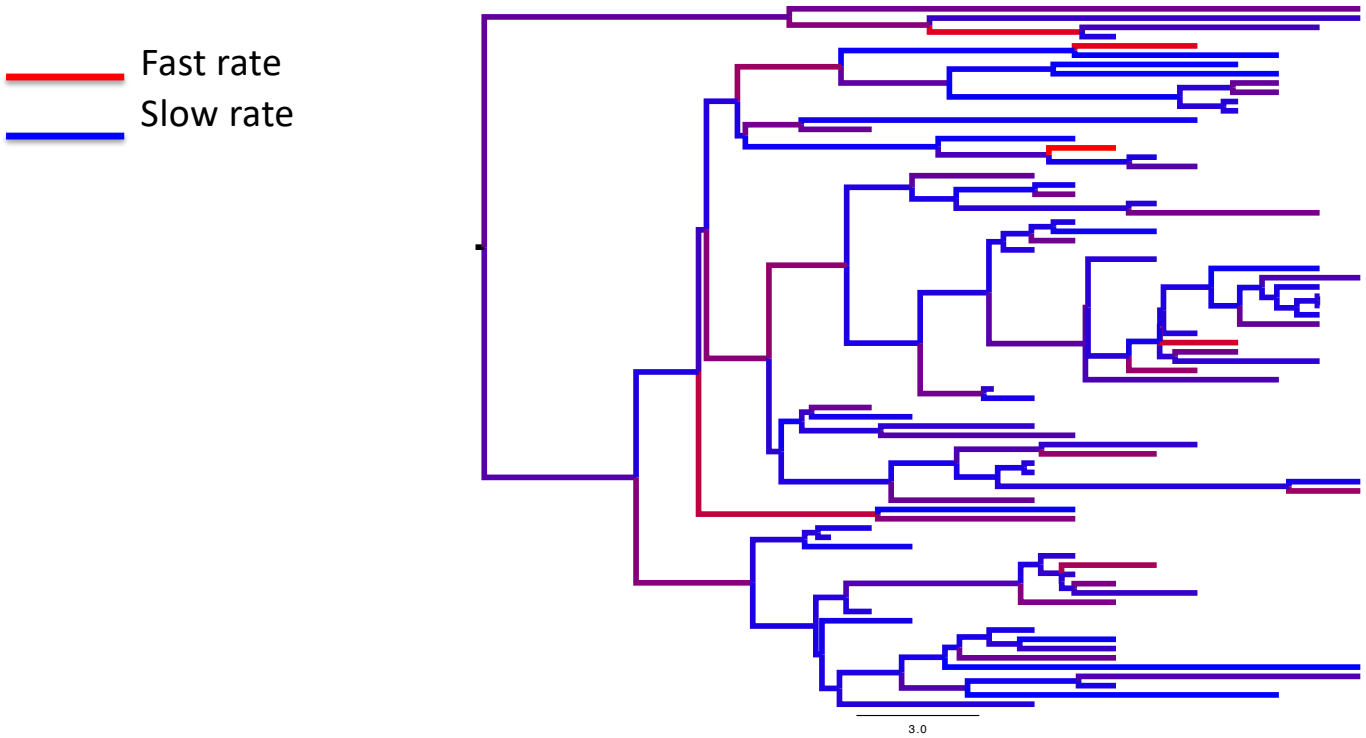
Current estimates: prior to 1924



Worobey et al. 2008, Nature

Evolution often not strictly clock-like

Relaxed clocks allow for rates to vary among branches (Drummond 2006, PLoS Biol)

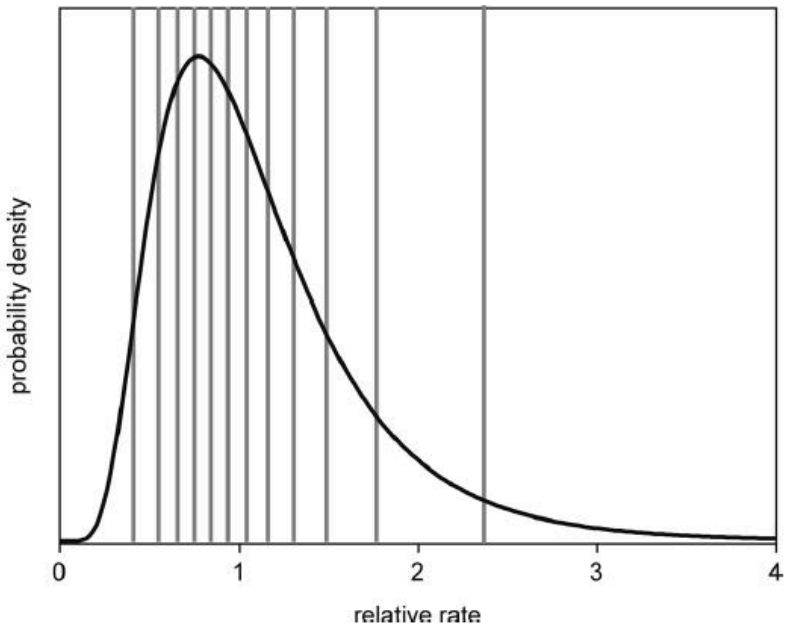


OPEN ACCESS Freely available online PLOS BIOLOGY

Relaxed Phylogenetics and Dating with Confidence

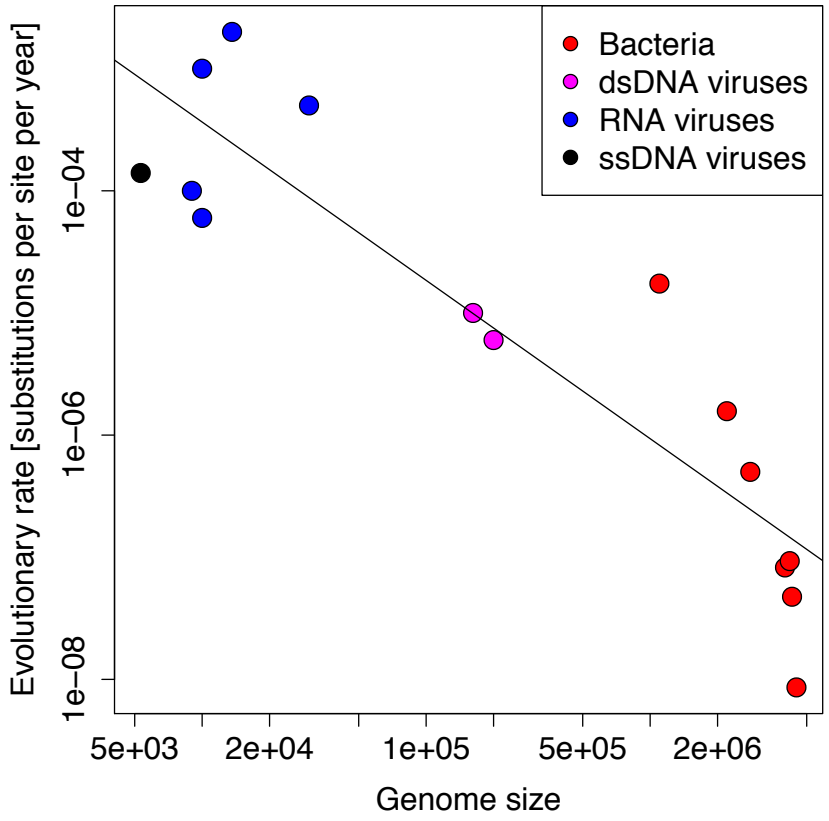
Alexei J. Drummond, Simon Y. W. Ho, Matthew J. Phillips, Andrew Rambaut*

Department of Zoology, University of Oxford, Oxford, United Kingdom



Relaxed molecular clock (uncorrelated): rate for each branch drawn randomly from distribution (lognormal or exponential)

Rate of evolution scales with genome size

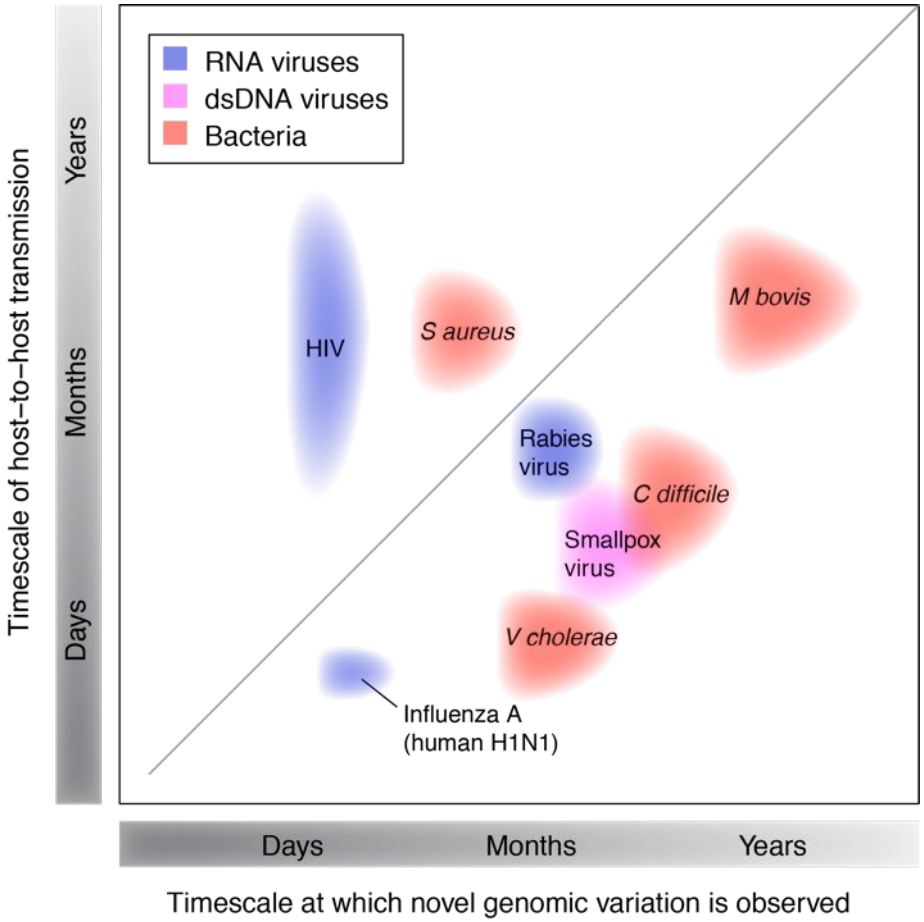


Evolutionary rate per genome much more similar than per site

Whole genome sequencing makes DNA viruses and bacteria measurably evolving

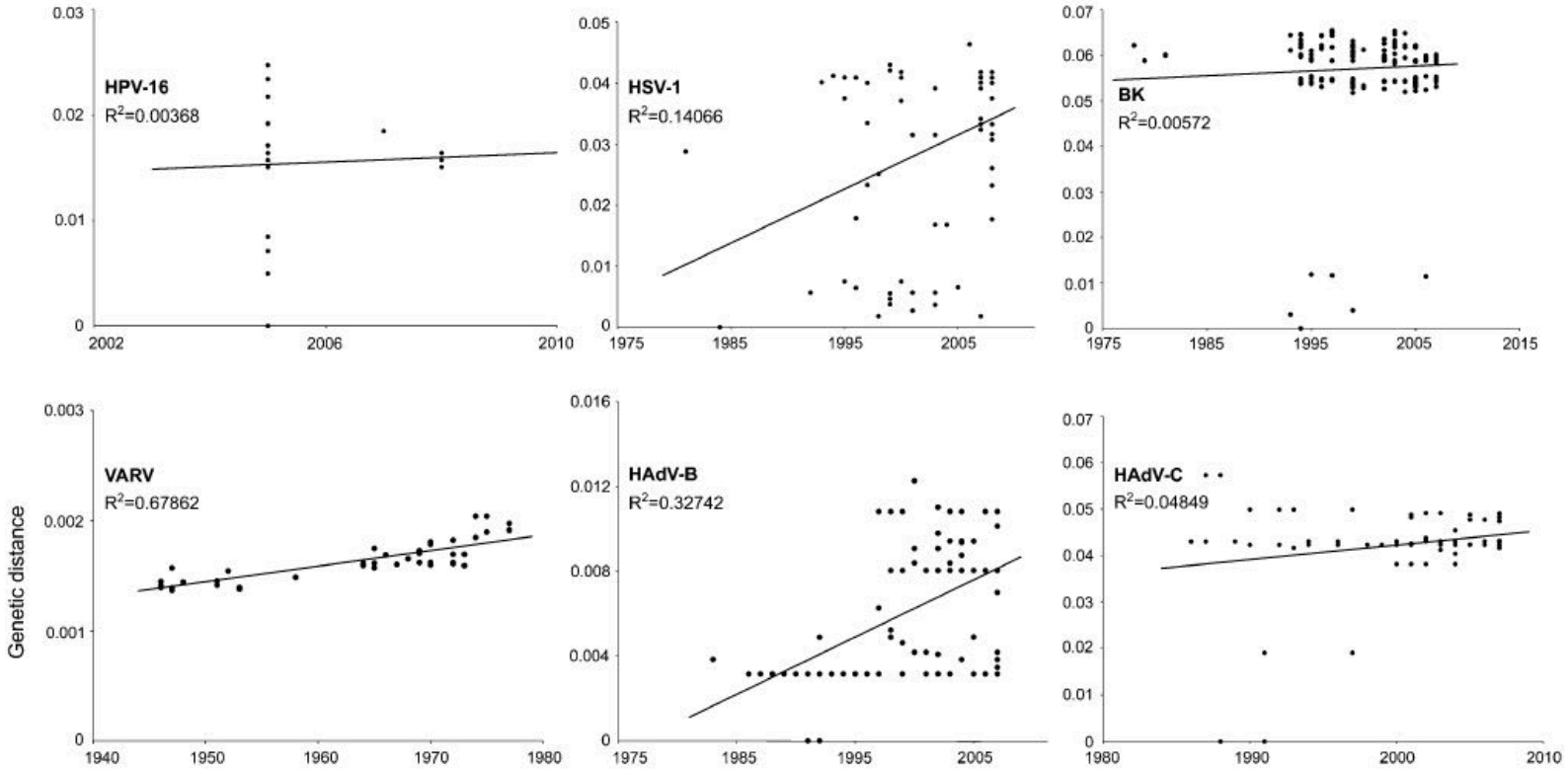
Biek et al. 2015, Trends Ecol Evol

Relative time scales of mutational and epidemiological events



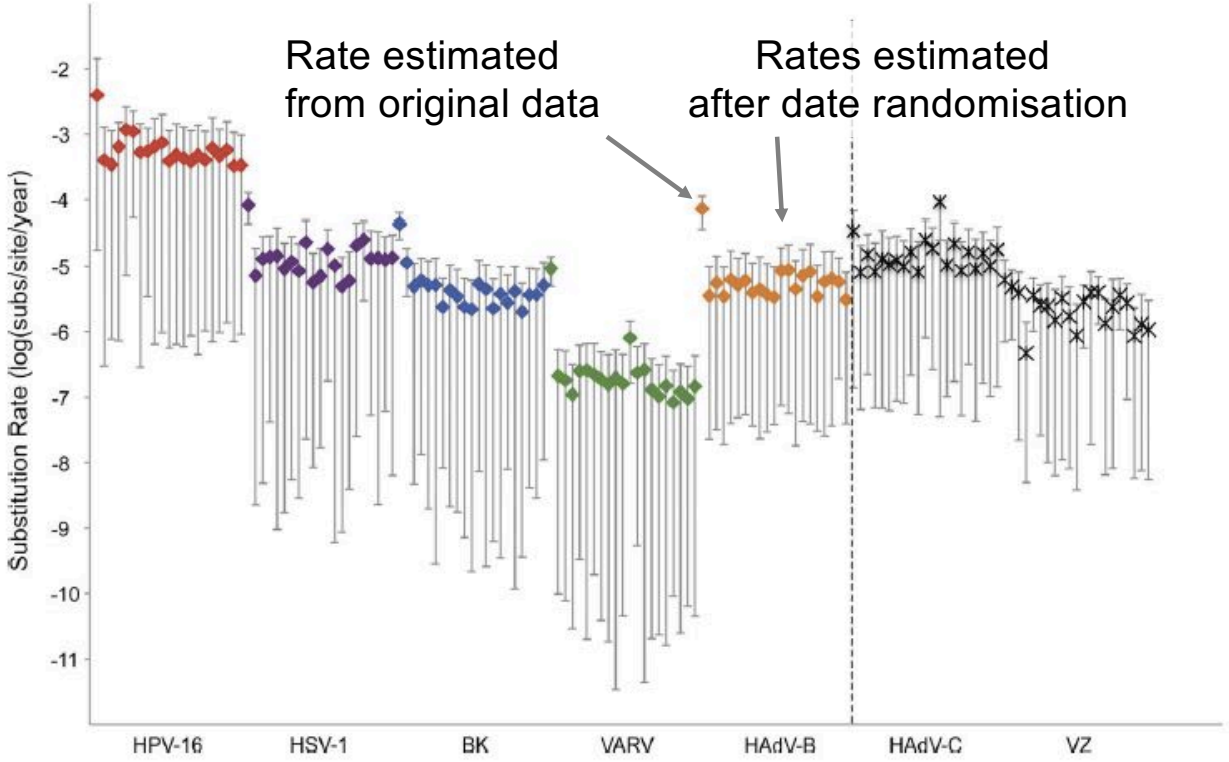
Biek et al. 2015, Trends Ecol Evol

Testing for measurable evolution in dsDNA viruses



Firth et al. 2010, Mol Biol Evol

Testing for measurable evolution in dsDNA viruses



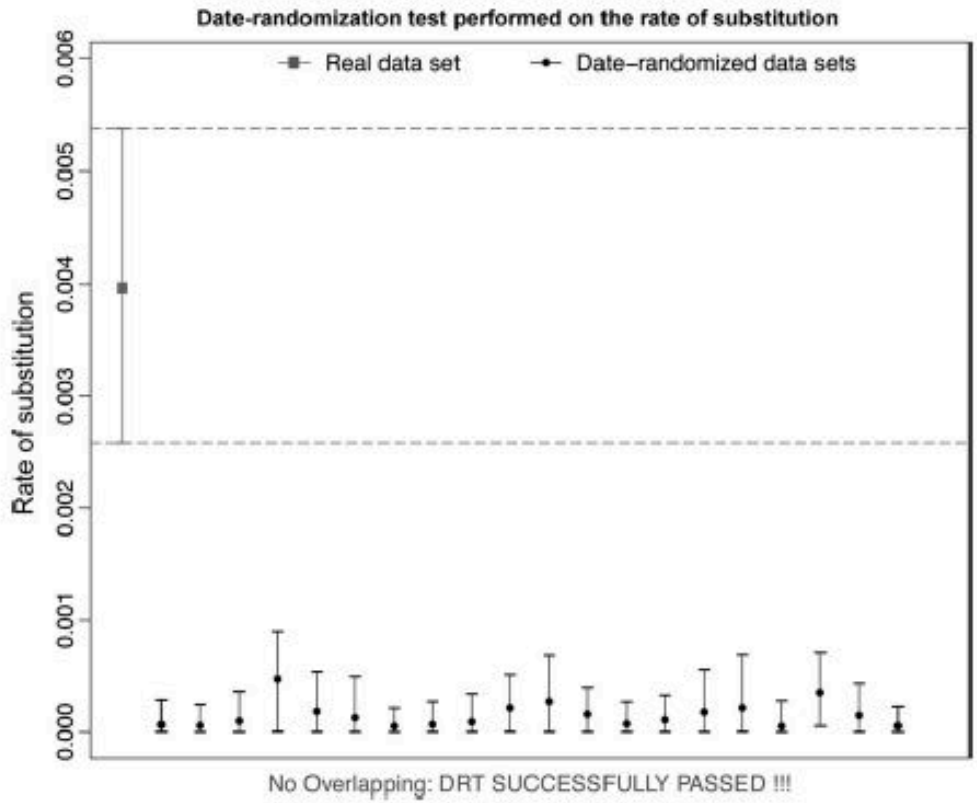
Firth et al. 2010, Mol Biol Evol

R package for date randomisation

MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) doi: 10.1111/1755-0998.12603

TIPDATINGBEAST: an R package to assist the implementation of phylogenetic tip-dating tests using BEAST

ADRIEN RIEUX* and CAMILO E. KHATCHIKIAN†‡
*CIRAD, UMR PVBMT, 97410 St Pierre, La Réunion, France, †Department of Biology, University of Pennsylvania, 433 S University Ave, Philadelphia, PA 19104, USA, ‡Department of Biological Sciences, The University of Texas at El Paso, 500 W. University Ave, Bioscience Research Bldg. Room 2.120, EL Paso TX 79968, USA



Virus origins and molecular clocks

- RNA viruses have been infecting vertebrates for probably millions of years

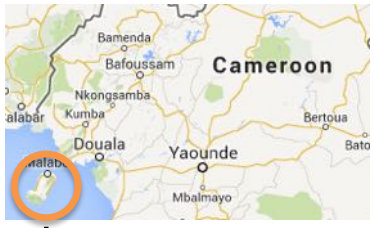
$\Rightarrow$ Some clear examples of co-evolution (e.g. foamy viruses, Switzer et al. 2005, Nature)

- But for most RNA viruses, molecular clocks would suggest common ancestors of a few thousand years (Holmes 2003, J Virol)

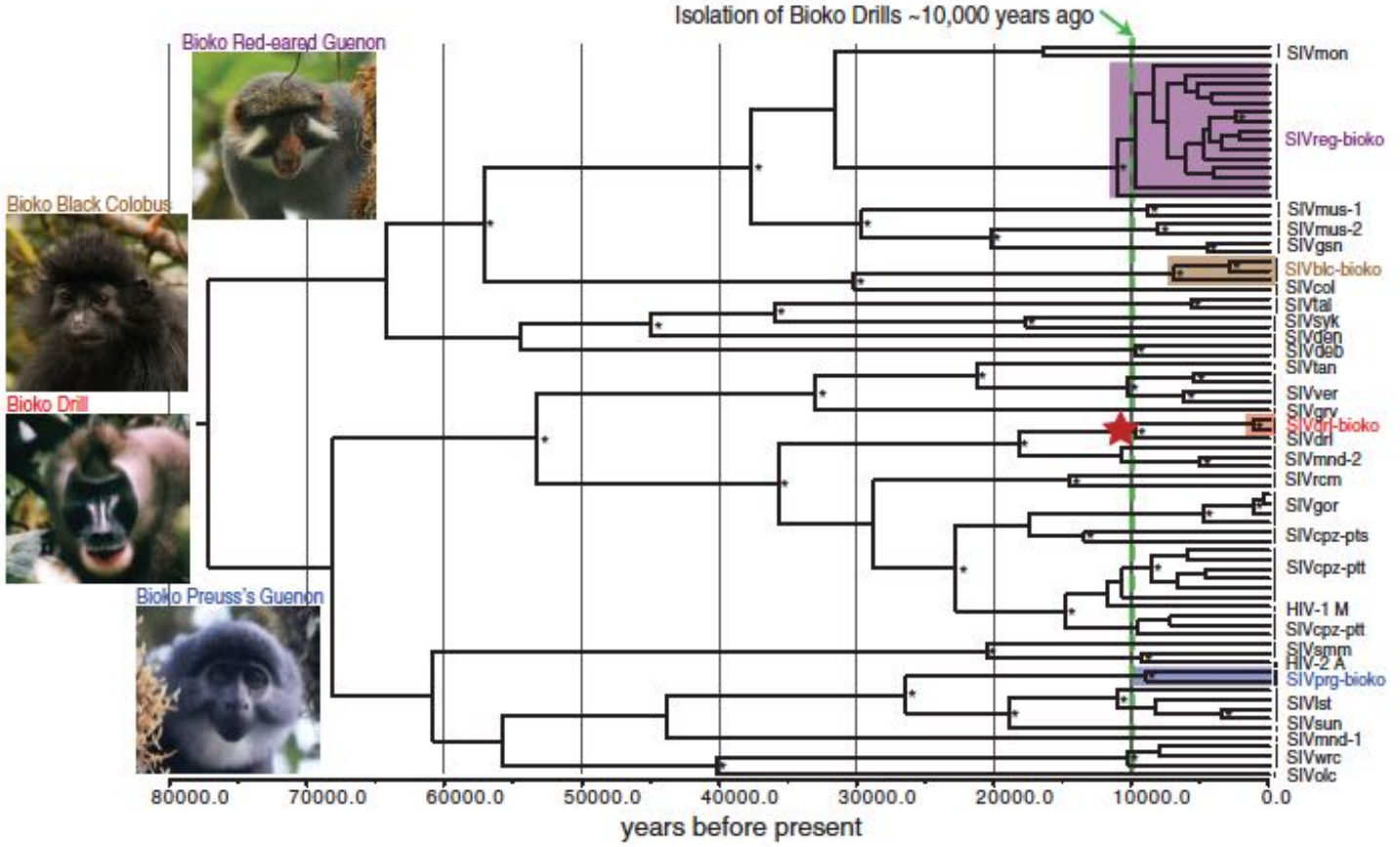
Clock rates estimated over time-spans of years or decades only valid on those time scales!

Improving rate estimates through node calibrations

SIV in primates on island of Bioko

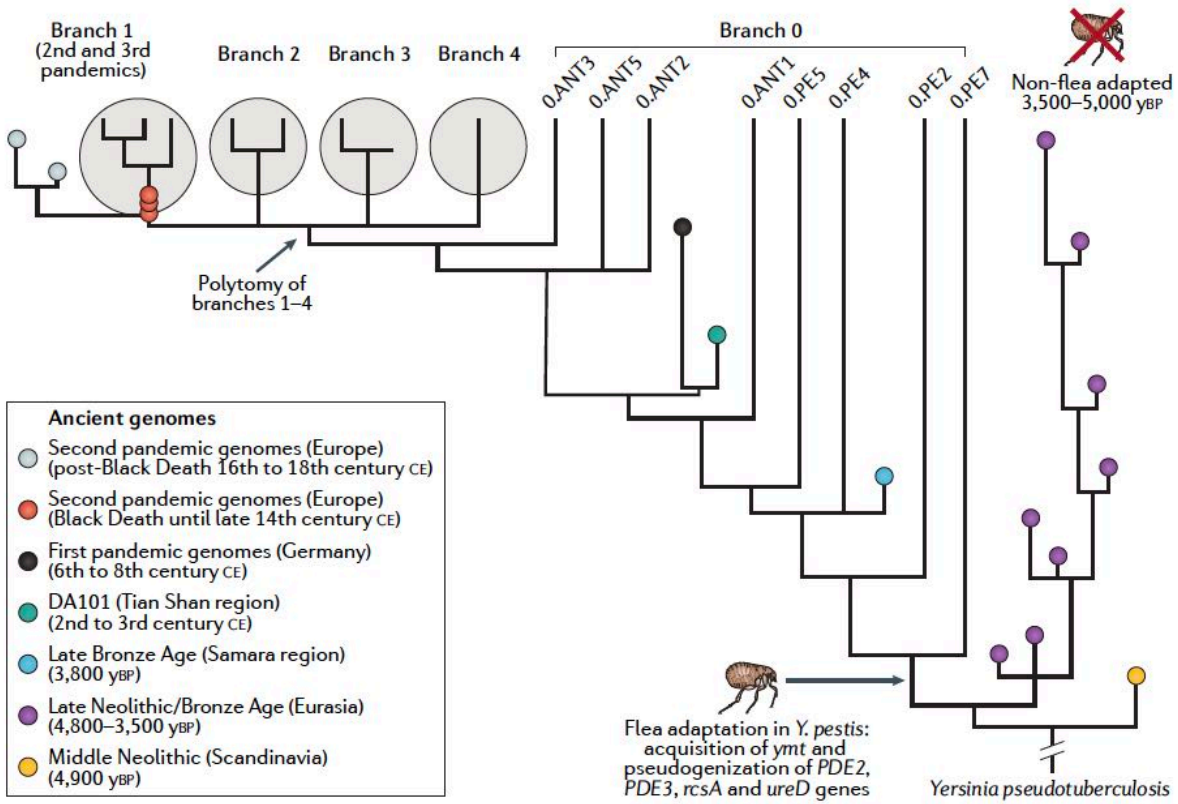


Bioko



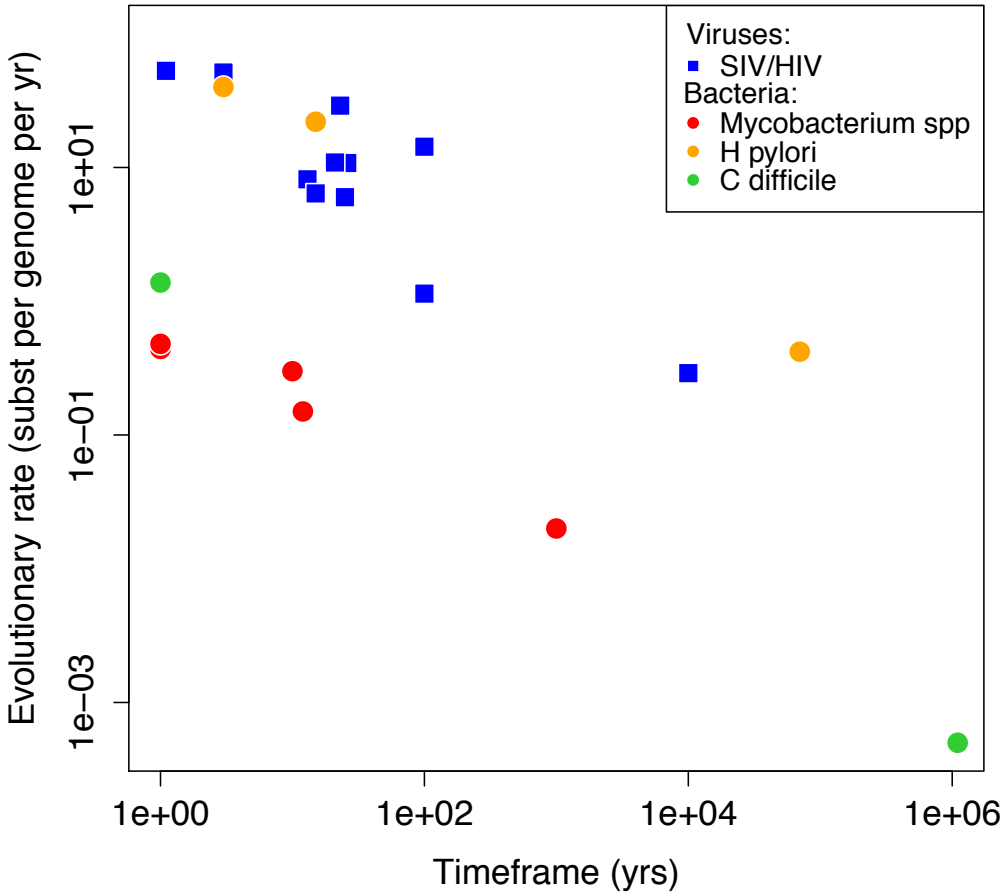
Worobey et al. 2010, Science

Ancient bacterial genomes



Spiyrou et al. 2019, Nature Rev Genetics

Ancient genomes reveal time-dependent decline in evolutionary rates

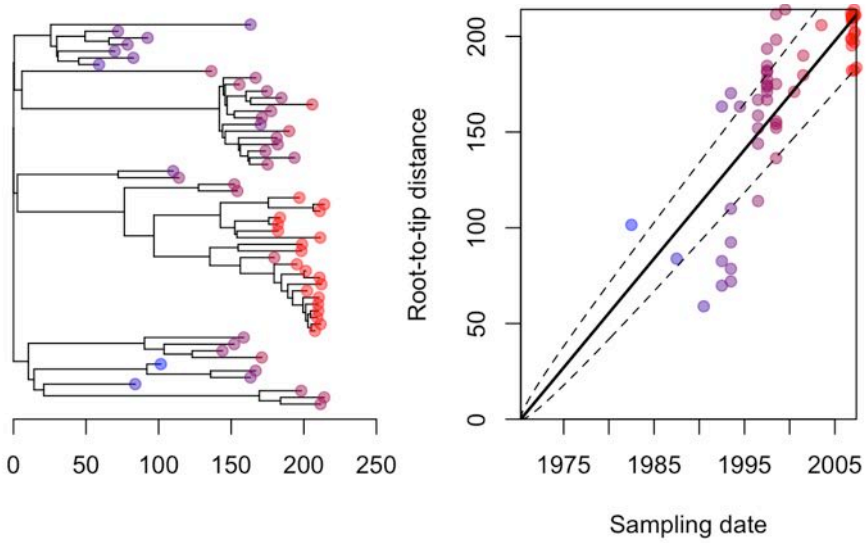


Estimated clock rate only valid for time scale of sampling!

Biek et al. 2015, Trends Ecol Evol

Fast molecular clock analyses of bacterial genomes

- BacDating – R package that implements Bayesian method to construct dated phylogenies for bacterial genomes (Didelot 2018, Nucleic Acids Research)
- Based on single genome so ignoring phylogenetic uncertainty



Program BEAST

Bayesian **E**volutionary **A**nalysis **S**ampling **T**rees

Program geared towards molecular clock analyses

Estimating phylogenies is not the focus: instead uses MCMC to average over tree space

Original developers:
Andrew Rambaut (U Edinburgh) – maintains version 1

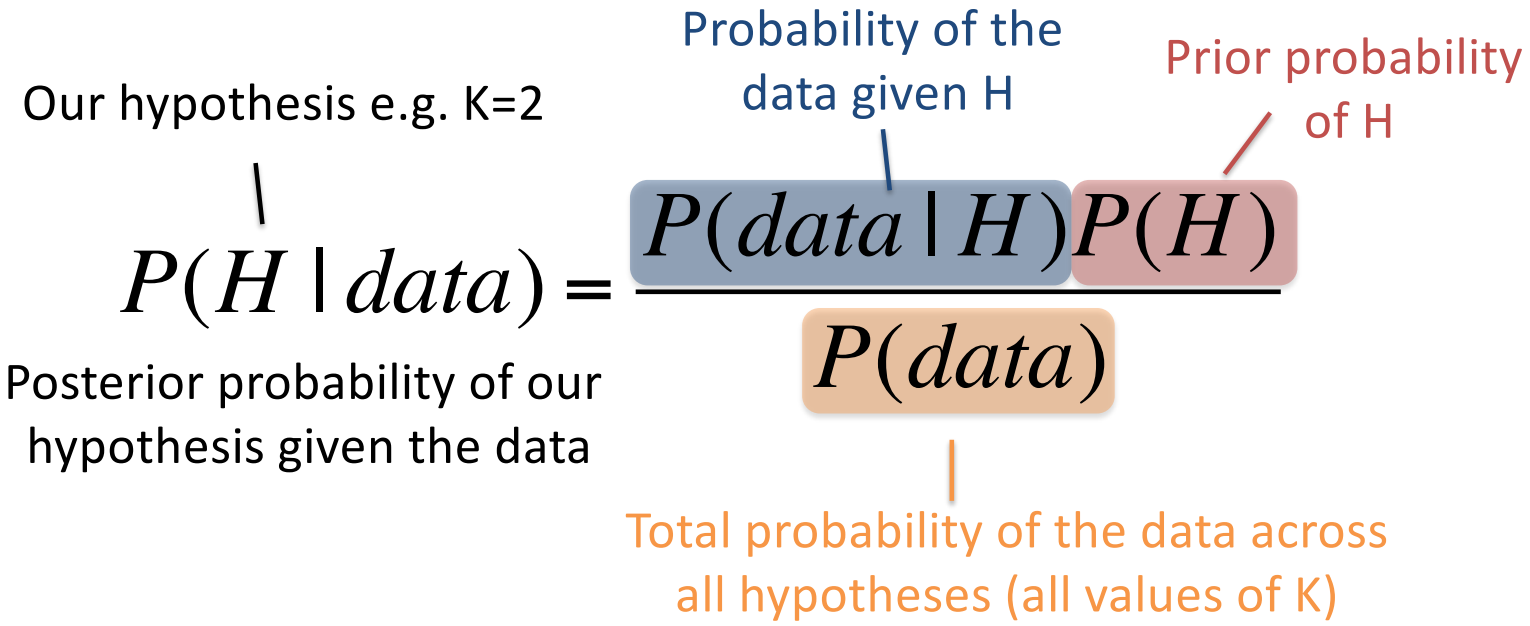
Alexei Drummond (U Auckland) – started BEAST2

Bayesian statistics

Reverend Thomas Bayes

1702-1761

Bayes rule in phylogenetic estimation

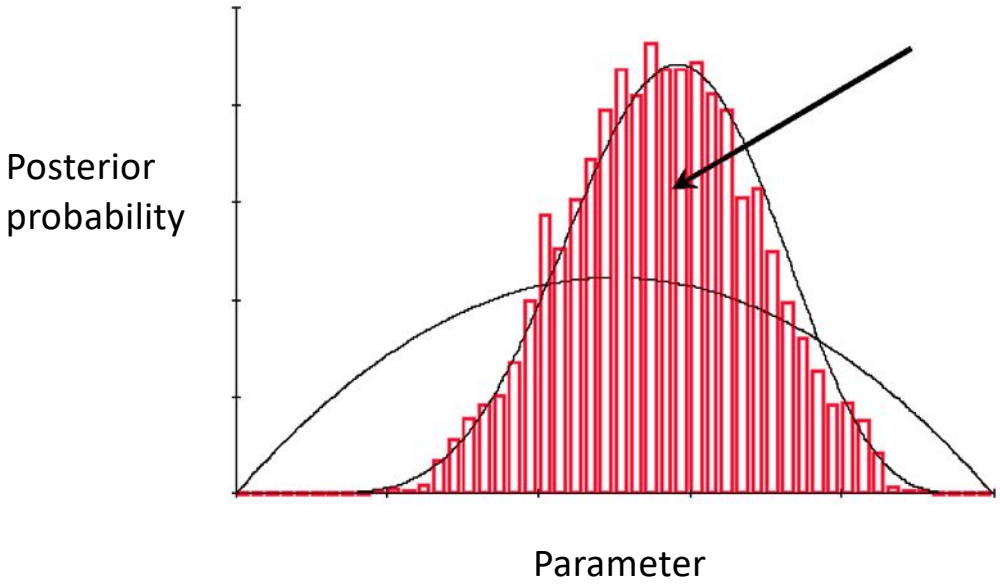


Multi-dimensional problem requires MCMC

- Computationally difficult to calculate posterior for many parameters simultaneously (e.g. tree topology, evolutionary substitution model parameters, etc)

$\Rightarrow$ Approximate posterior using Markov Chain Monte Carlo (MCMC) method

Markov chain Monte Carlo (MCMC)

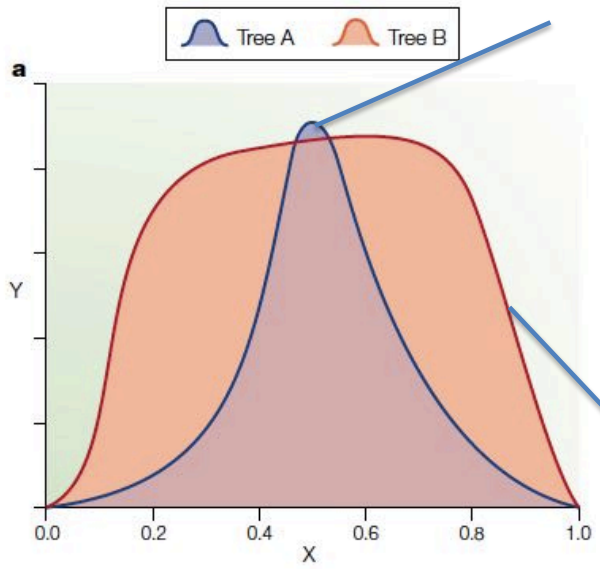


instead of deriving this distribution through integration, sample from a simulated distribution that is expected to be the posterior distribution

From Mark de Been and Rob Willems

Bayesian vs Maximum Likelihood

Likelihood (ML) or posterior probability (Bayesian)



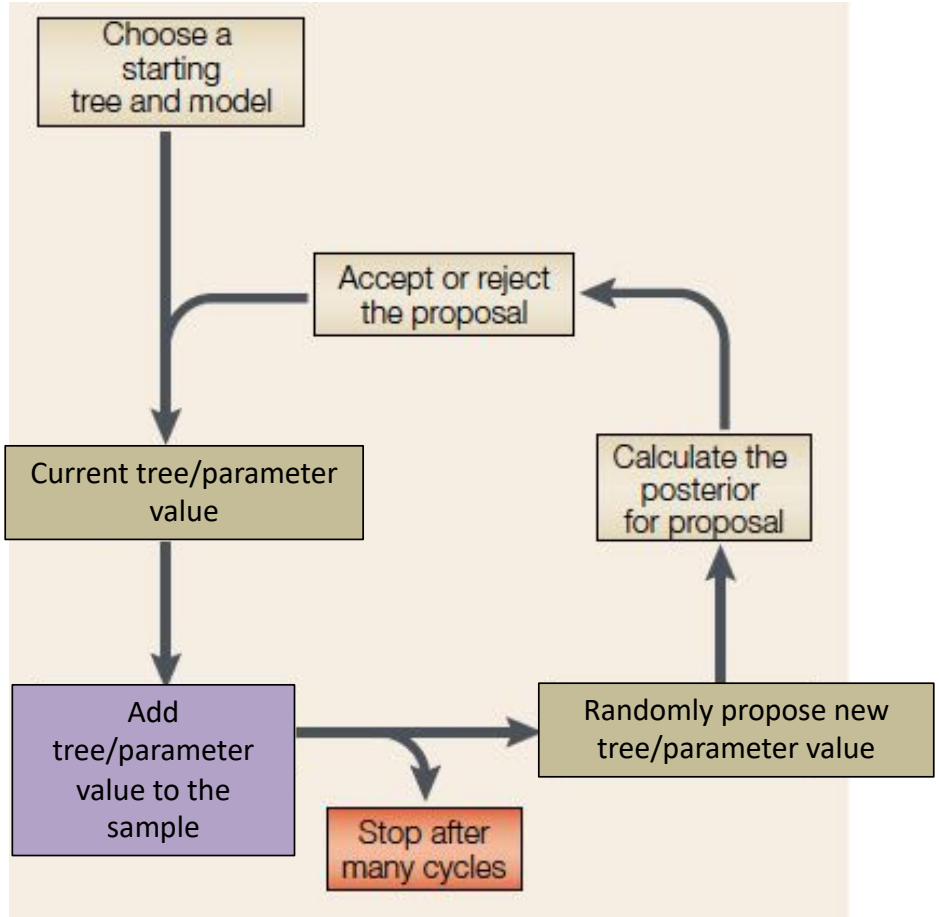
Parameter such as transition/transversion ratio

Joint estimation of parameter and tree (ML) favours this tree (has the higher likelihood peak)

Marginal estimation of parameter and tree (Bayesian) favours this tree (area under the curve is larger)

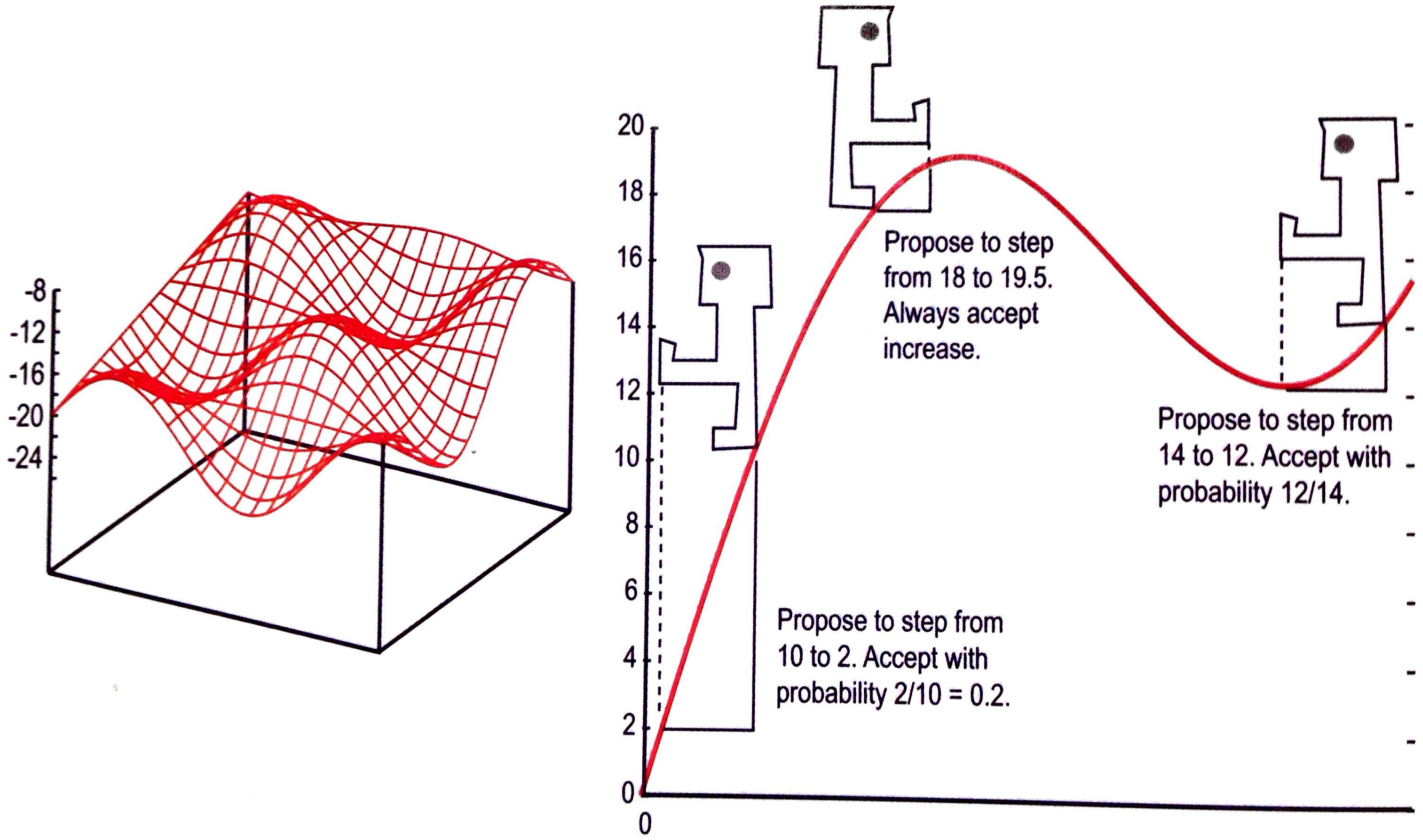
Holder and Lewis et al 2003, Nature Reviews Genetics

Estimation of parameters and trees in Bayesian phylogenetics



Modified from: Holder and Lewis et al 2003, Nature Reviews Genetics

MCMC robot



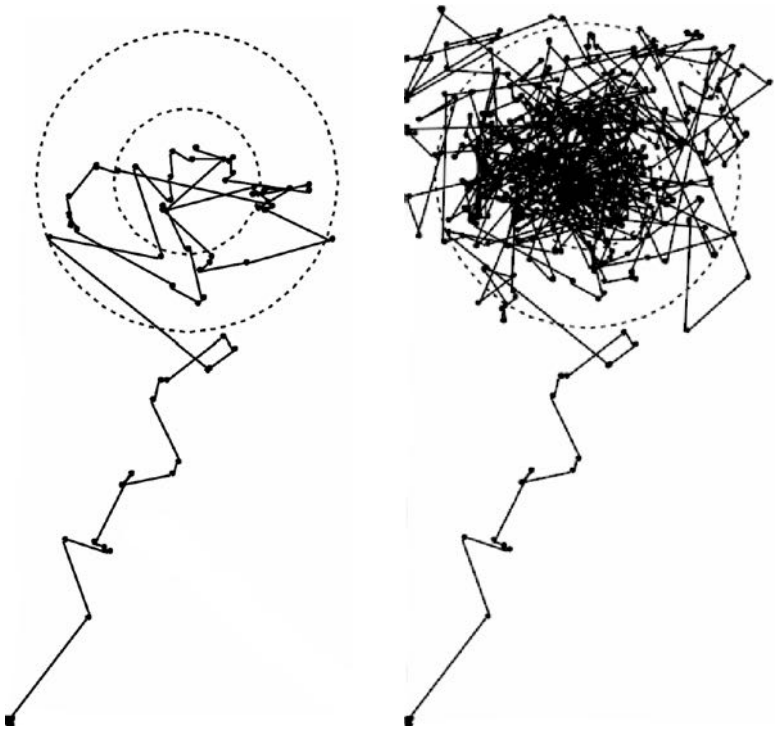
From Drummond and Bouckaert 2015

‘Burn-in’

takes some time for the robot (i.e. chain) to reach a hill (area of high posterior probability)

⇒not representative of the posterior probability

⇒duration of ‘burn-in’ varies considerably with dataset



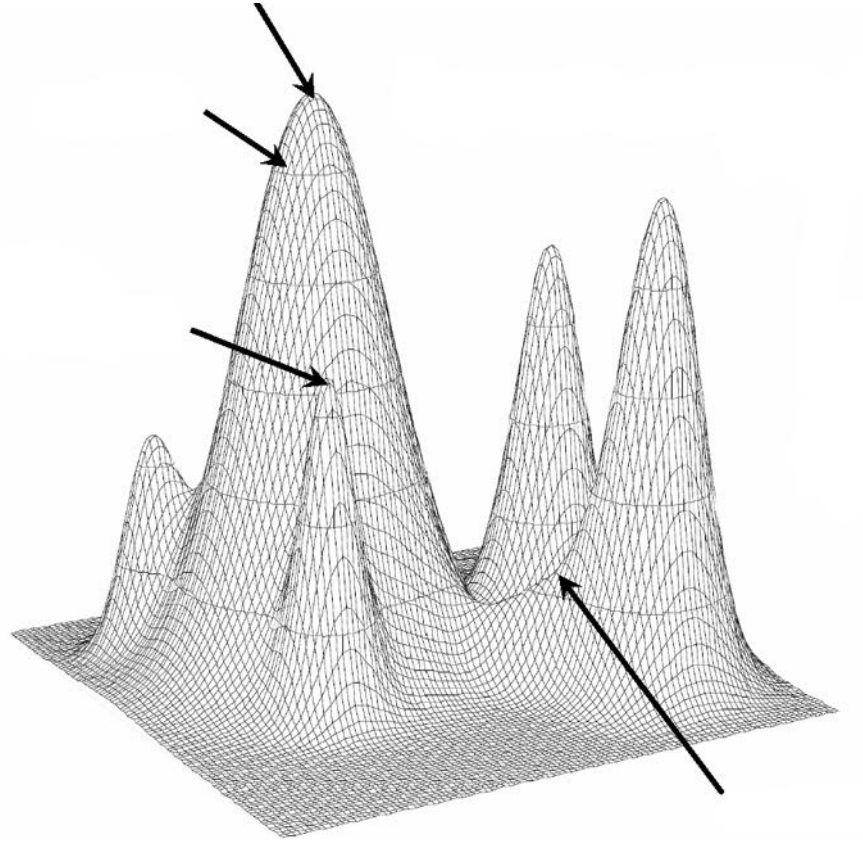
From Mark de Been and Rob Willems

Problems with MCMC

"poor mixing"

MCMC chain may spend long periods of time stuck in one place

chain may only find local optima



From Mark de Been and Rob Willems

Molecular clock summary

- Measurably evolving populations accumulate substitutions over the time span of sampling
- Means that molecular clock can be estimated and used to put time-scale on phylogeny
- Used to be RNA viruses only, with genomic data also applicable to bacteria and DNA viruses
- Pitfalls:
  - insufficient temporal signal (should check!)
  - rate depends on time scale of sampling