SISMID
Module 14: Evolutionary Dynamics and Molecular Epidemiology of Viruses

Instructors: Julia A. Palacios and Nicola Müeller

TAs: Nídia Trovão and Joëlle Barido-Sottani

Logistics

- Zoom sessions are recorded and will be available after the session.

- Other instructors will be available in slack for questions and discussions during zoom sessions.

- For this lecture, Nicola is available in slack for questions.

https://juliapalacios.github.io/SISMID_EvolutionaryDynamics/

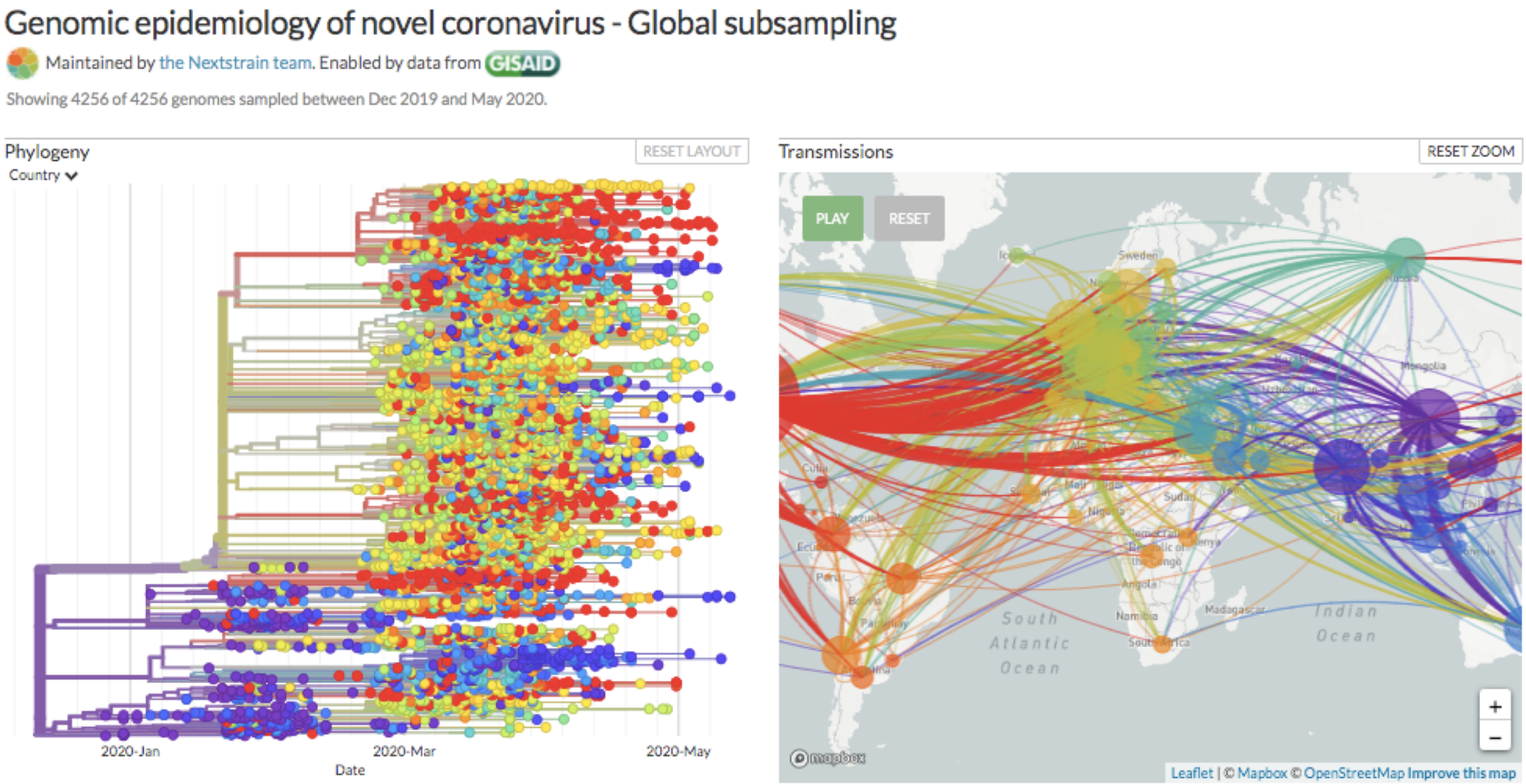
Evolutionary dynamics and molecular epidemiology of viruses

The goal is to:

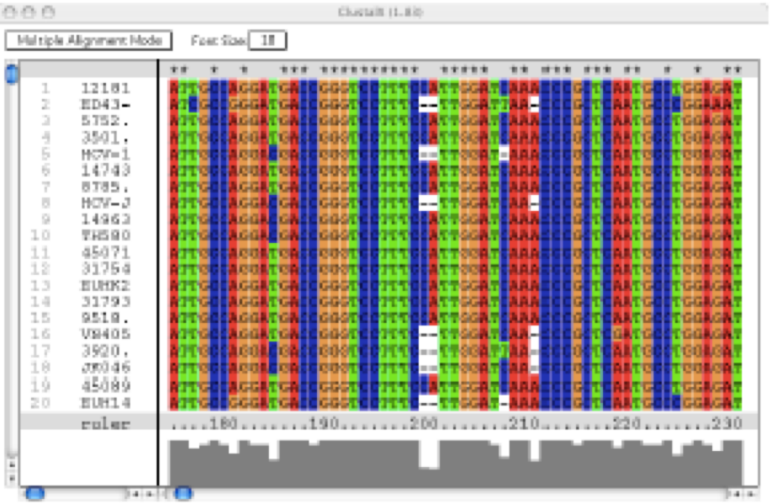
- Understand patterns of transmission and spread (effective population size)
- Estimate the rate of evolution / mutation rate
- Compare evolution across pathogens
- Understand the sources of molecular variation (mutation, selection, recombination)
- Surveillance

Molecular epidemiology and **phylodynamics** of infectious diseases aim to study infectious disease behavior through a combination of evolutionary, epidemiological and immunological processes from molecular variation [HG09].

Global spread of SARS-CoV-2



Observed data



- **Biological sequences** (DNA, RNA, protein) contain information about their underlying evolutionary processes.
- Molecular sequences from different organisms are **not independent** because they share evolutionary history.
- The central concept is a **genealogy**: a bifurcating tree that depicts the **ancestral relationships** of the samples.

Estimation of genealogies and phylogenies has allowed …

PNAS

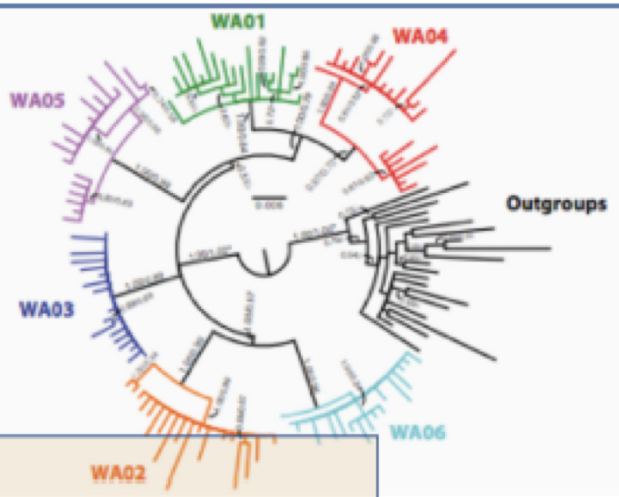
Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences

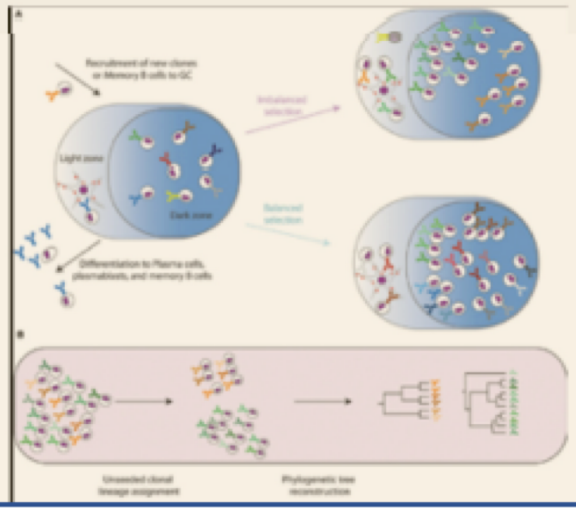
Diane I. Scaduto[a,b], Jeremy M. Brown[c,1], Wade C. Haaland[a,b], Derrick J. Zwickl[c,2], David M. Hillis[c,3], and Michael L. Metzker[a,b,d]

[a]Human Genome Sequencing Center, [b]Department of Molecular and Human Genetics, and [c]Cell and Molecular Biology Program, Baylor College of Me Houston, TX 77030; and [c]Section of Integrative Biology and Center for Computational Biology and Bioinformatics, University of Texas, Austin, TX 7

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2008.

Contributed by David M. Hillis, October 20, 2010 (sent for review September 23, 2010)





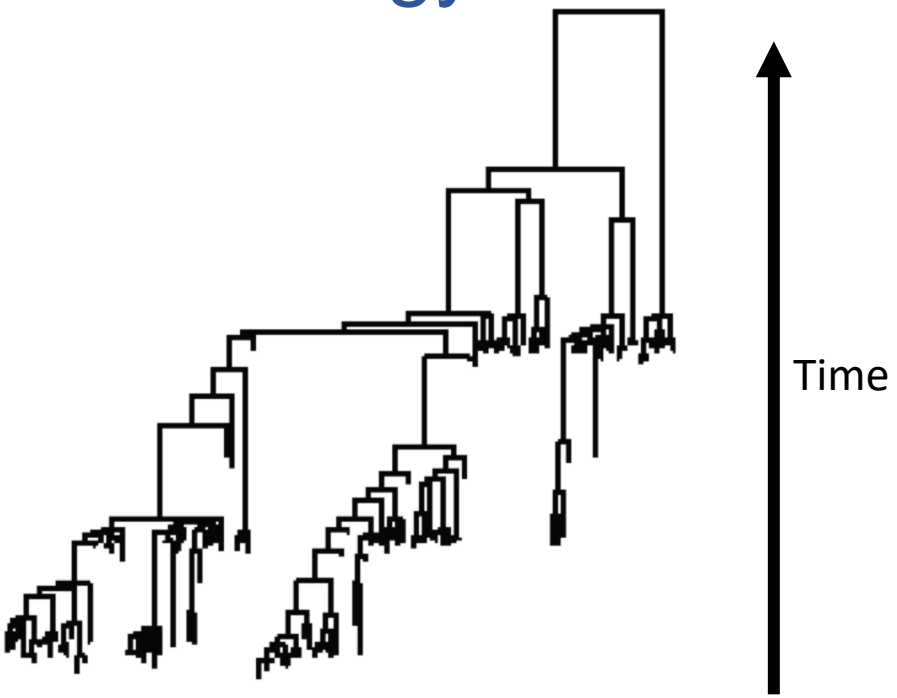
Tracing Antibody Repertoire Evolution by Systems Phylogeny

Alexander Dimitri Yermanos[1,2], Andreas Kevin Dounas[3], Tanja Stadler[1], Annette Oxenius[2] and Sai T. Reddy[3]

[1]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland, [2]Department of Biology, Institute of Microbiology, ETH Zurich, Zurich, Switzerland, [3]Department of Chemistry and Applied Biosciences, ETH Zurich, Zurich, Switzerland

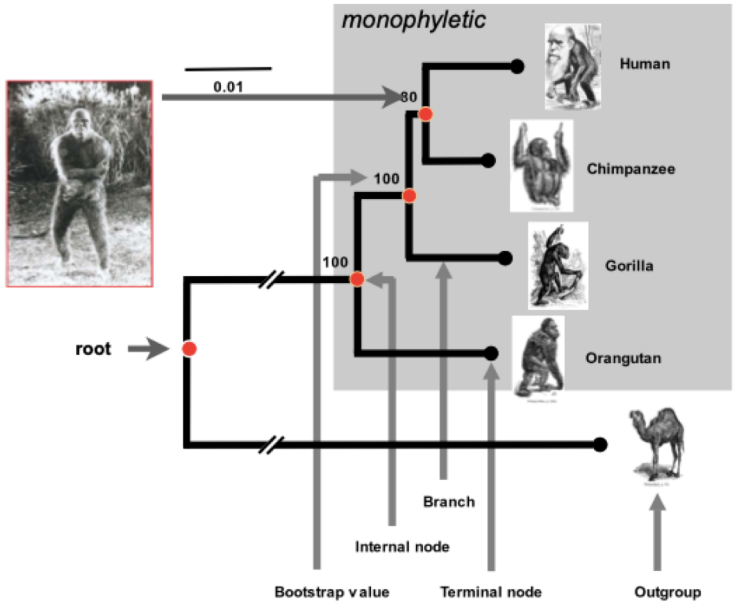
What is a genealogy?

Genealogy



- Tips correspond to individuals
- Internal nodes are ranked
- Branch lengths are in the same scale
- Samples are time stamped (tips)

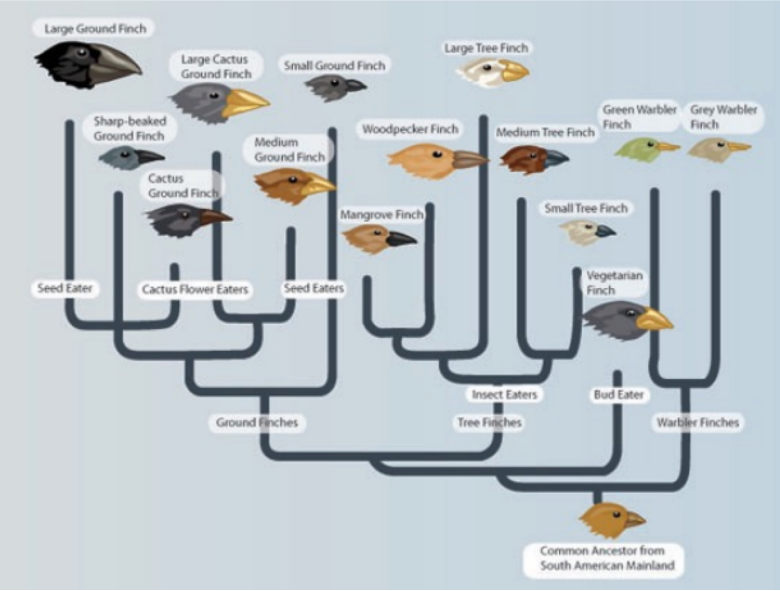
Phylogeny



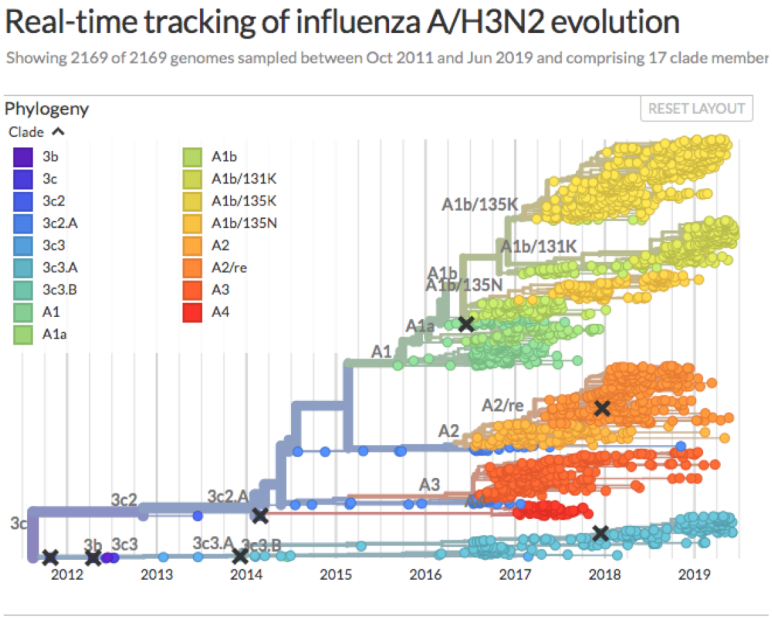
- Tips correspond to species
- Usually internal nodes are not ranked
- Branch lengths are in different scales
- Unrooted trees are commonly analyzed

Phylogenetics, phylodynamics and population genetics

- **Phylogenetics** is the study of the evolutionary history of species. It seeks to determine the "family tree".
  - Understanding selection
  - Evidence for coevolution
  - Pathways of trait evolution

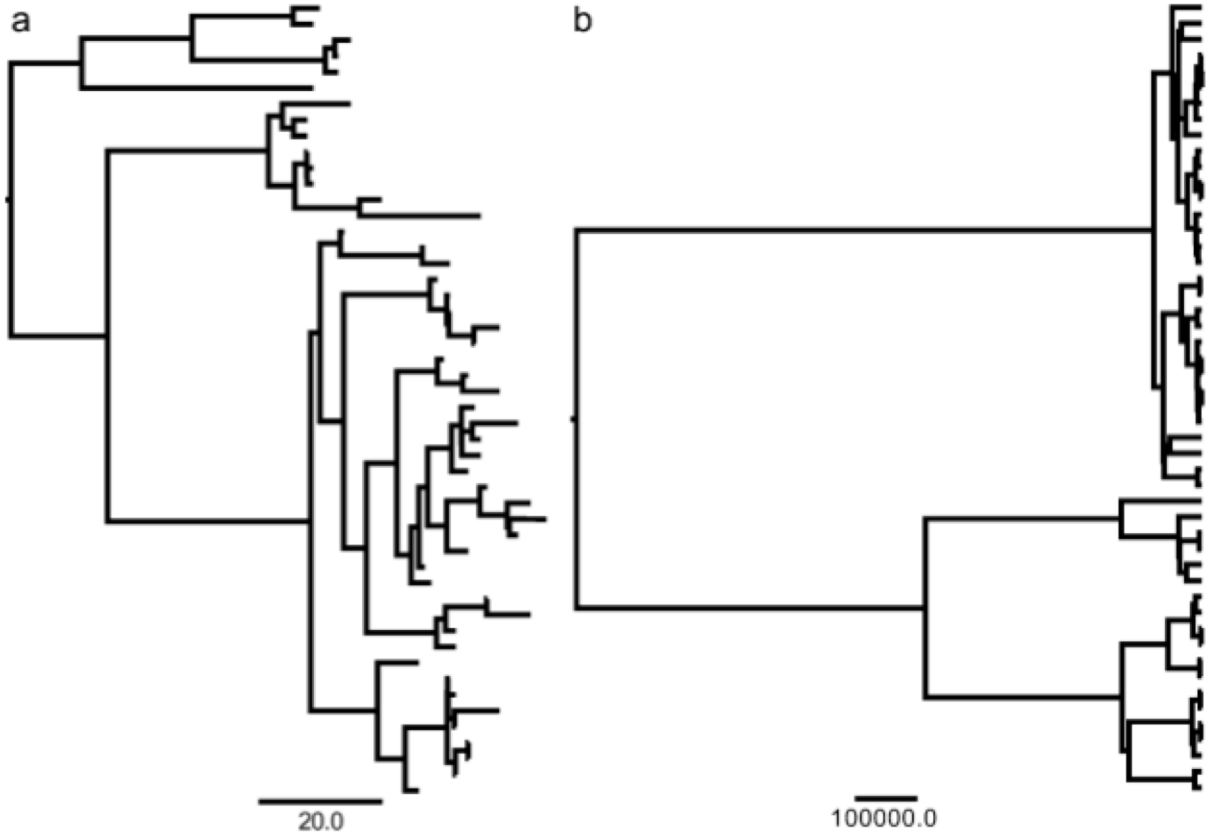


- **Phylodynamics**
  - Attempts to enhance understanding of infectious disease dynamics using pathogen phylogenies



Phylodynamics and population genetics

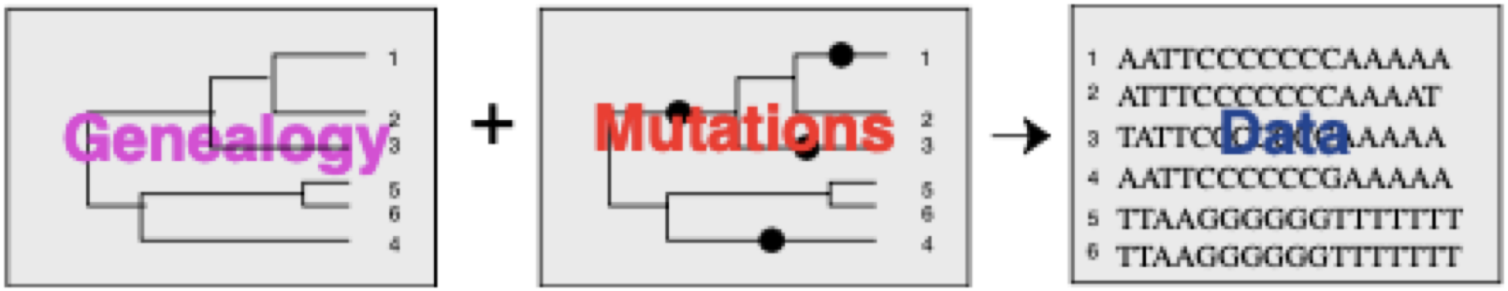
- **For rapidly evolving organisms**
- **For slowly evolving organisms**



10^{-4} subs/site/year

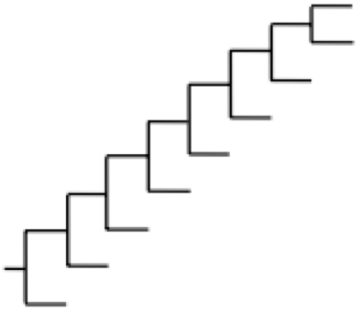
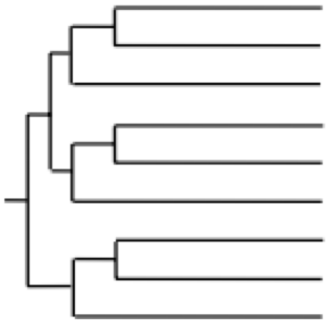
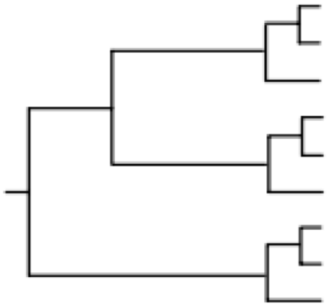
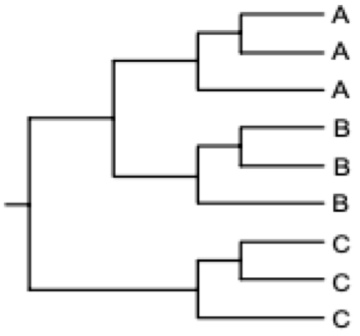
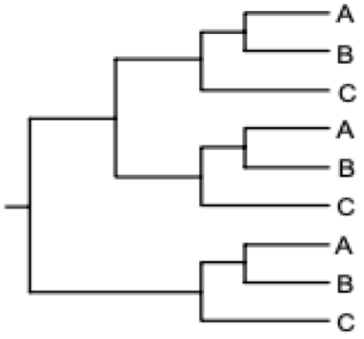
10^{-8} subs/site/year

Statistical Phylogenetics seeks to infer genealogies from molecular data

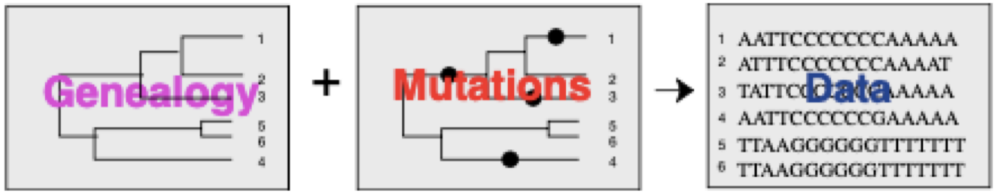


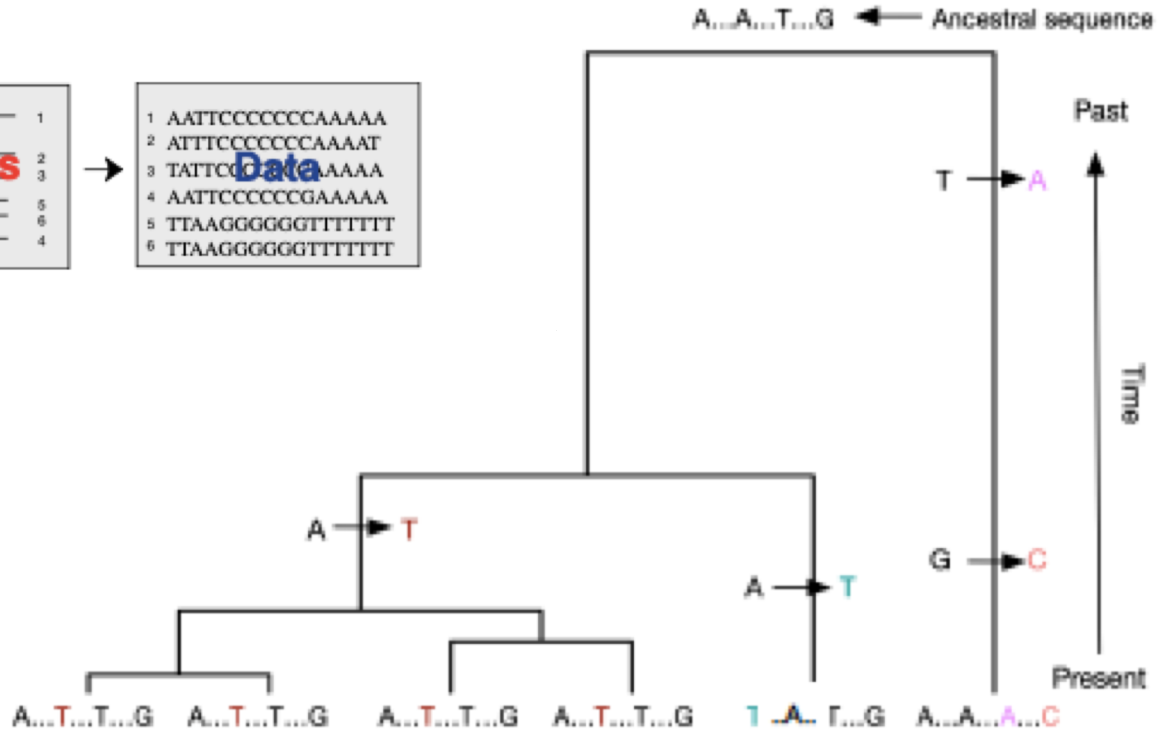
- Genealogies inform about past evolutionary history.
  - Ancestry
  - Signatures of selection
  - Population structure
  - Population history

Phylodynamic Patterns

	Continual Immune Selection	Weak/No Immune Selection	
		Population dynamics	Spatial dynamics
Idealised Phylogeny Shapes		*Population growth*  *Population decline* 	*Strong spatial structure*  *Weak spatial structure* 
Examples	Human influenza A within-host HIV	among-host HIV among-host HCV	Measles Rabies, Dengue

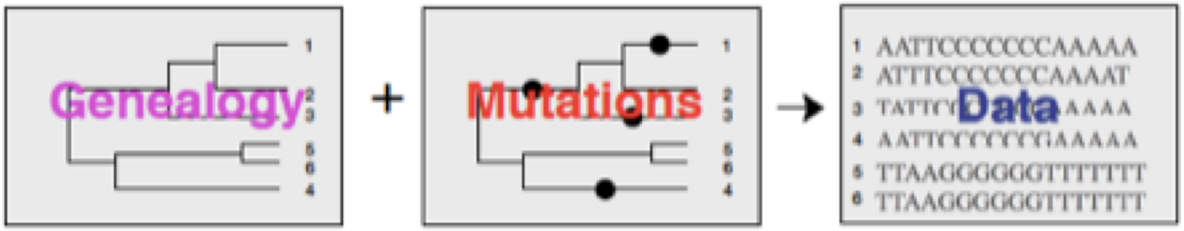
A process of substitutions superimposed on the genealogy **generates** observed sequences at the tips of the genealogy





Statistical Phylogenetics

Goal: Estimate genealogy/phylogeny

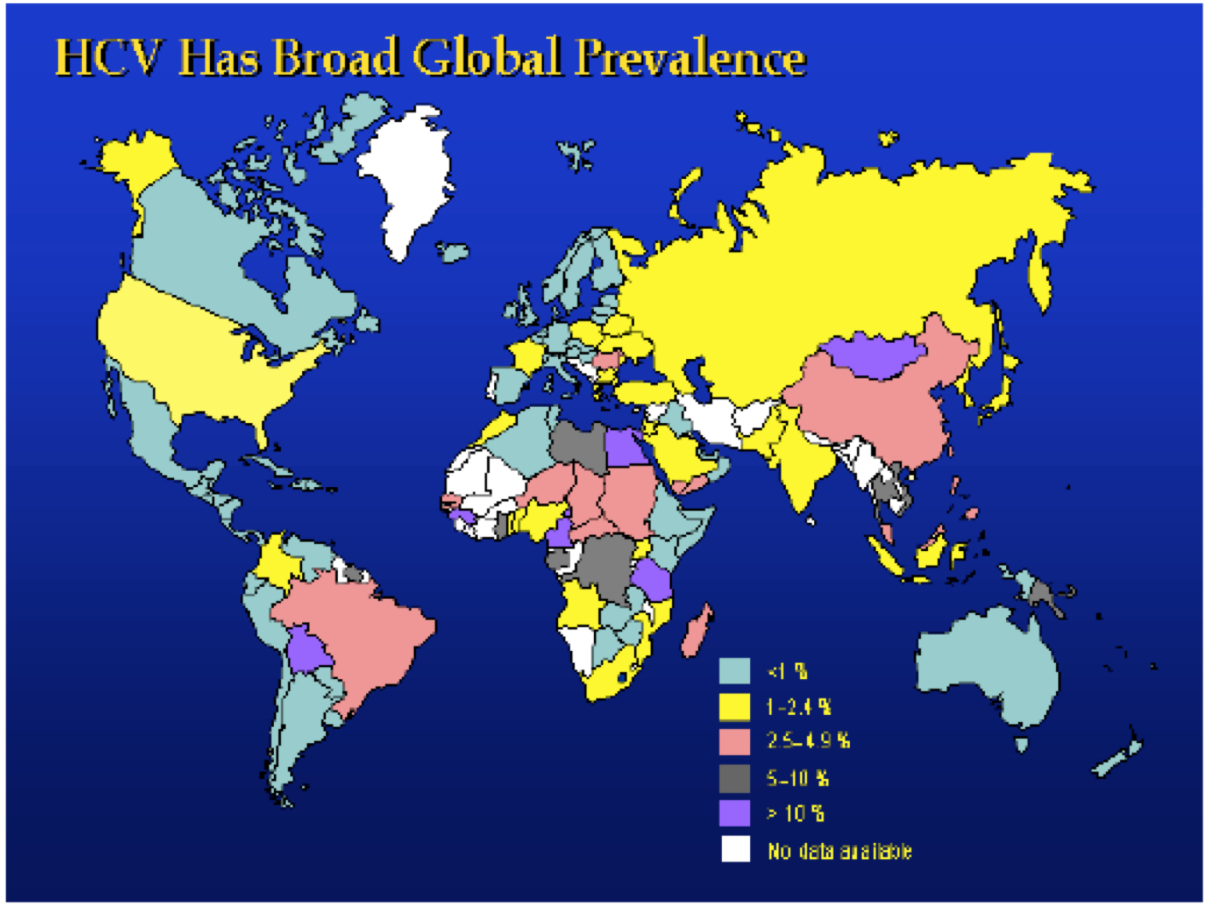


Phylodynamics

Goal: Estimate effective population size $N_e(t)$ from DNA sequences

Coalescent Process

Example 1: Hepatitis C in Egypt



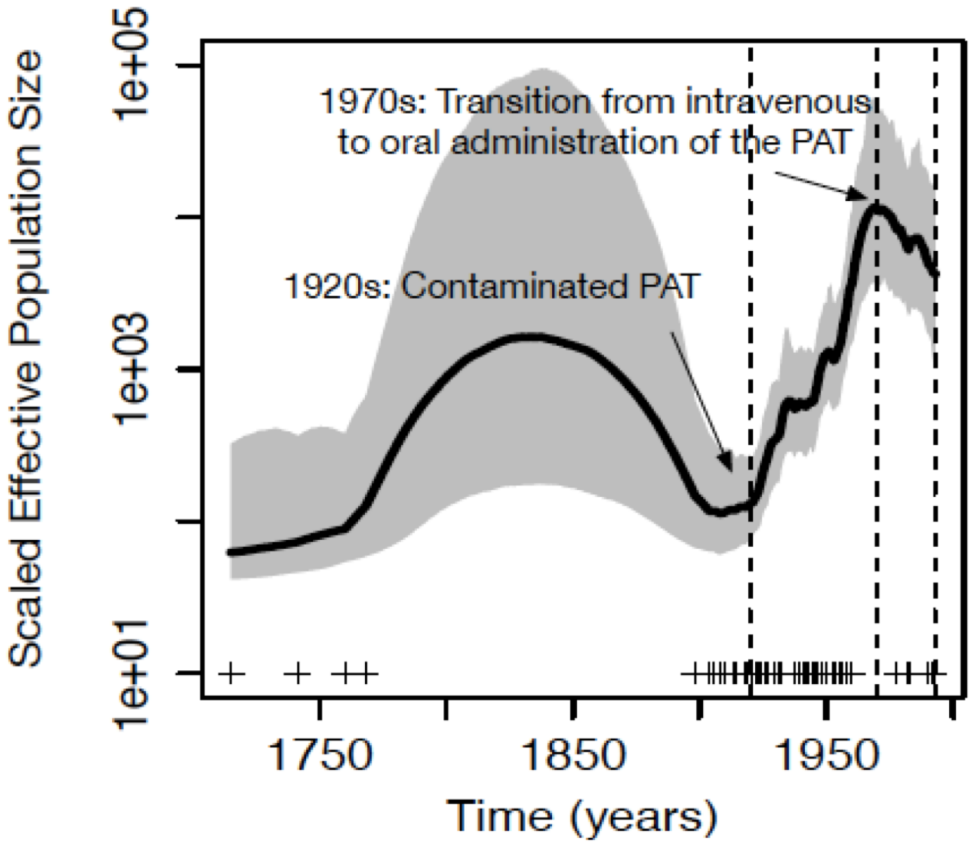
Prevalence of HCV - WHO 1999

- Identified in 1989
- Spread by blood to blood contact
- ≈3% of infected population worldwide
- 8,000 - 10,000 deaths per year in the USA
- Egypt has the highest prevalence

Example 1: Hepatitis C in Egypt

- 62 samples in 1993 from the E1 gene (411bp)
- Parenteral antischistosomal therapy (PAT) was practiced from 1920s to 1980s
- In the 1970s started a transition from the intravenous to the oral administration of the PAT

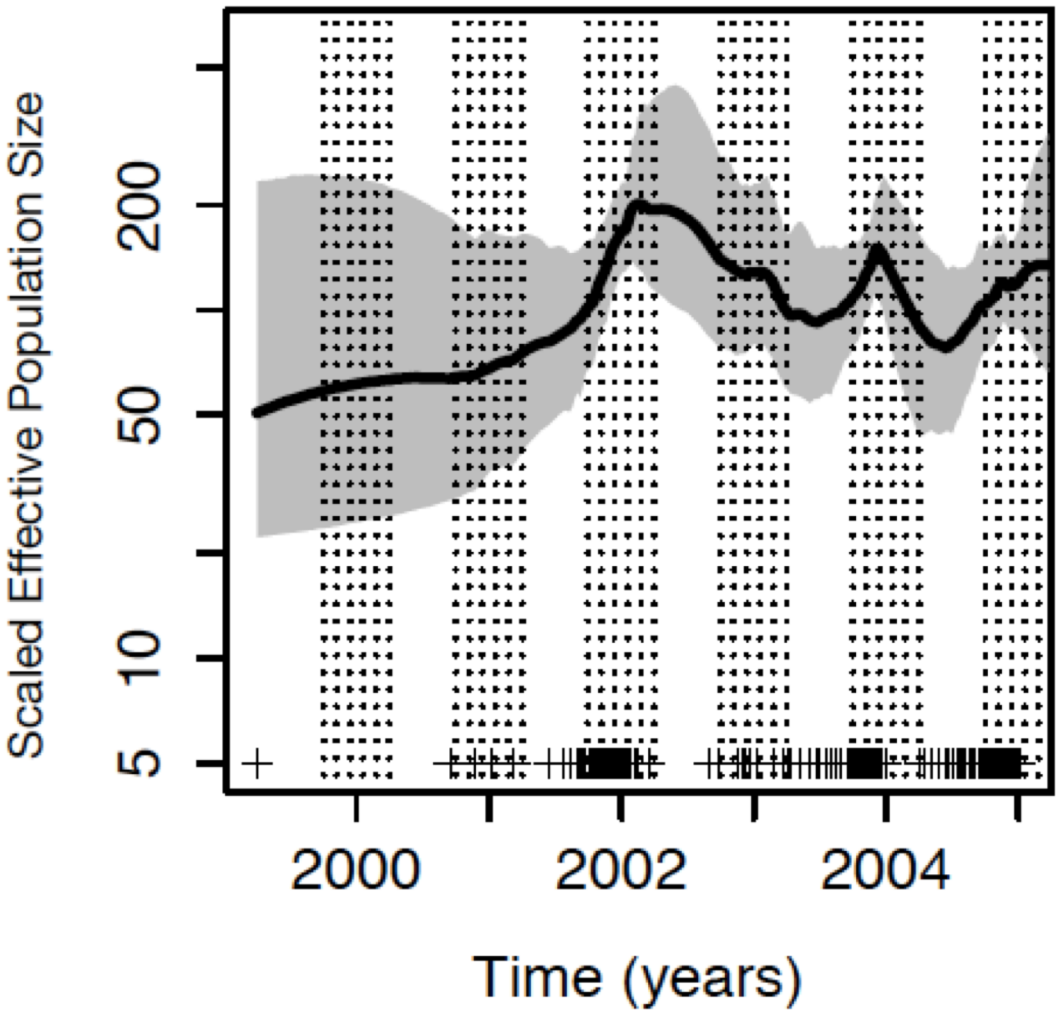
Example 1: Hepatitis C in Egypt



[Palacios and Minin, Biometrics 2013]

- 62 samples in 1993 from the E1 gene (411bp)
- Parenteral antischistosomal therapy (PAT) was practiced from 1920s to 1980s
- In the 1970s started a transition from the intravenous to the oral administration of the PAT

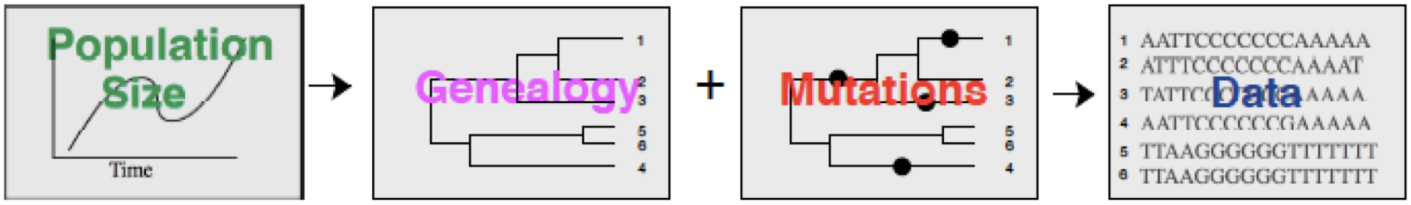
Example 2: Influenza in NY



- Human Influenza A in N.Y.
- 288 sequences from 2001-2005 from HA gene

[Palacios and Minin, Biometrics 2013]

Bayesian Evolutionary Analysis by Sampling Trees



$N_e(t)$ $\quad$ $\mathbf{G}$ $\quad$ $\mathbf{Q}$ $\quad$ $\mathbf{Y}$

$$P(N_e(t), \mathbf{G}, \mathbf{Q}, \tau \mid \mathbf{Y}) \propto \underbrace{P(\mathbf{Y} \mid \mathbf{G}, \mathbf{Q})}_{\text{Likelihood}} \underbrace{P(\mathbf{G} \mid N_e(t))}_{\text{Coalescent prior}} P(\mathbf{Q}) \underbrace{P(N_e(t) \mid \tau)}_{\log \mathrm{GP}(0, \mathbf{C}(\tau))} P(\tau)$$

Posterior

Likelihood

Coalescent prior
Birth-death prior

log GP$(0, \mathbf{C}(\tau))$
Piece-wise constant (deterministic)
Parametric prior

Frequentist vs Bayesian Inference

- Frequentist
  - Probability is interpreted as long run frequency.
  - The goal is to create procedures with long run guarantees.
  - Procedures are random while parameters are fixed and unknown
- Bayesian
  - Probability is interpreted as a measure of subjective degree of belief
  - Everything is regarded as random
  - Goal is to quantify and analyze degrees of belief

Larry Wasserman –All of Statistics

Bayesian Inference

- We begin with a *prior* belief about the values of the parameters $\theta \in \Theta$ of the model.

$$\pi(\theta) \tag{1}$$

  This express your belief about θ before you have seen the data.
- The sampling distribution (or likelihood) has a known functional form: $L(X_1 \ldots, X_n \mid \theta)$.
- Applying Bayes' rule, we get the following posterior distribution

$$P(\theta \mid X_1, \ldots, X_n) = \frac{L(X_1, \ldots, X_n \mid \theta)\pi(\theta)}{\int_{\theta \in \Theta} L(X_1, \ldots, X_n \mid \theta)\pi(\theta) d\theta} \tag{2}$$

Bayesian Inference

$$\pi(\boldsymbol{\theta}) \tag{3}$$

$$L(X_1, \ldots, X_n \mid \boldsymbol{\theta}) \tag{4}$$

If one is **philosophically Bayesian**, then the interpretation is the following: “Given my prior beliefs about the unknown parameters, my assumptions about the sampling model, and the data I have observed, my beliefs about the unknown parameters are now expressed by the posterior, the conditional distribution of parameters given data"

$$P(\boldsymbol{\theta} \mid X_1, \ldots, X_n) \tag{5}$$

Example: Poisson-Gamma

- Suppose your observation(s) is(are) a realization from a Poisson distribution with parameter λ =1
- You don't know that λ =1
- You have a prior belief that λ may behave as a Gamma(.1,1)

$$\mathrm{P}(\theta \mid x_1, \ldots, x_n) = \frac{\mathrm{P}(x_1,\ldots,x_n|\theta)\mathrm{P}(\theta)}{\mathrm{P}(x_1,\ldots,x_n)}$$

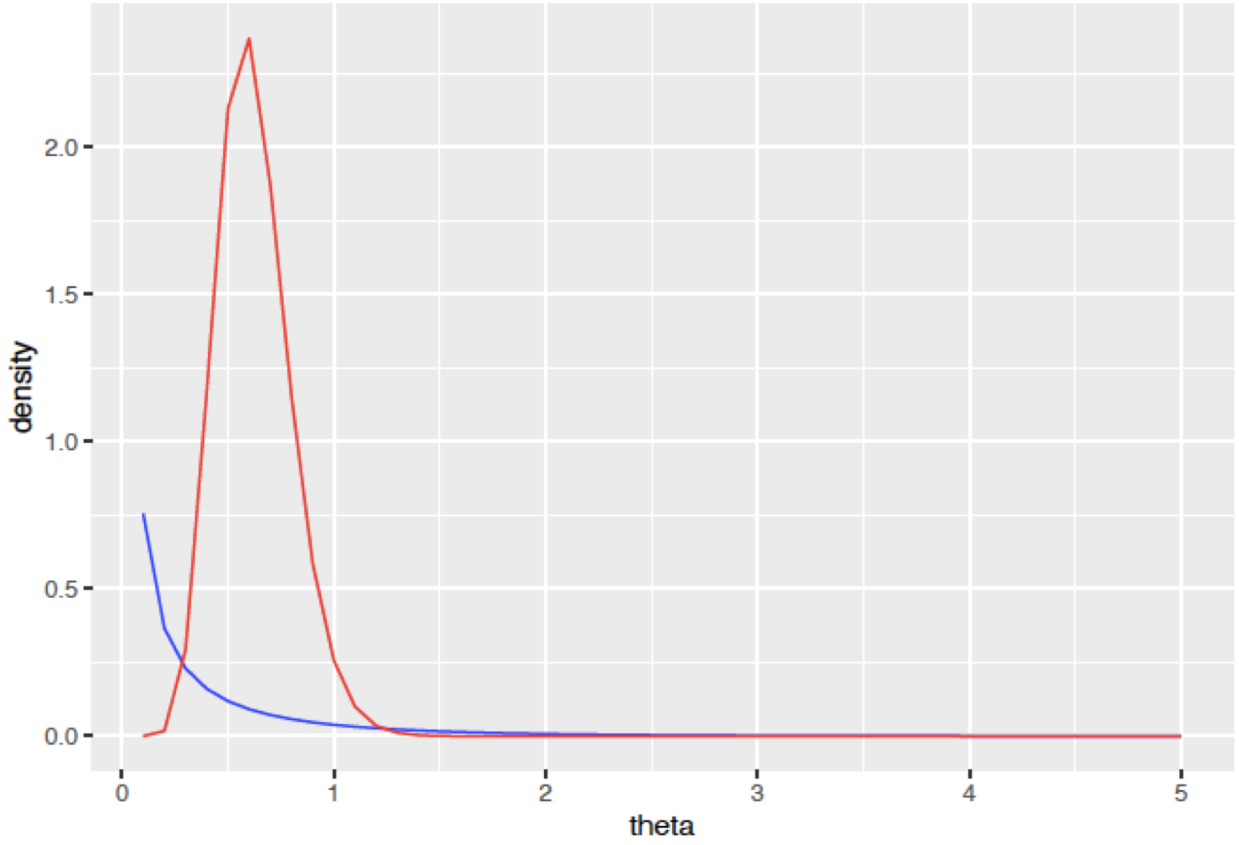
$$\mathrm{P}(\theta \mid x_1, \ldots, x_n) = \frac{\theta^{\sum_i^n x_i} e^{-\theta} (\prod_{i=1}^n x_i!)^{-1} \theta^{\alpha-1} e^{-\theta/\beta} (\Gamma(\alpha)\beta^\alpha)^{-1}}{\mathrm{P}(x_1,\ldots,x_n)}$$

$$\mathrm{Gamma}(\textstyle\sum_{i=1}^n x_i + \alpha, (n + 1/\beta)^{-1})$$

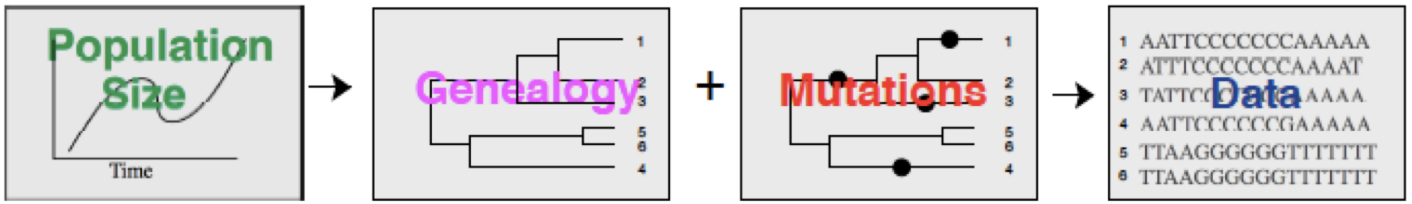
Example: Poisson-Gamma

```
n<-20
y<-rpois(n,1) #true theta=1
```

```
library("ggplot2")
x<-seq(0.1,5,by=.1)
prior<-dgamma(x,.1,1)
posterior<-dgamma(x,sum(y)+.1,1+n)
df<-data.frame(x=x,prior=prior,posterior=posterior)
ggplot() +
  geom_line(data = df, aes(x = x, y = prior), color = "blue") +
  geom_line(data = df, aes(x = x, y = posterior), color = "red")+ xlab('theta') +
  ylab('density')
```



Bayesian Evolutionary Analysis by Sampling Trees



$N_e(t)$ $\quad$ $\mathbf{G}$ $\quad$ $\mathbf{Q}$ $\quad$ $\mathbf{Y}$

$$P(N_e(t), \mathbf{G}, \mathbf{Q}, \tau \mid \mathbf{Y}) \propto \underbrace{P(\mathbf{Y} \mid \mathbf{G}, \mathbf{Q})}_{\text{Likelihood}} \underbrace{P(\mathbf{G} \mid N_e(t))}_{\text{Coalescent prior}} P(\mathbf{Q}) \underbrace{P(N_e(t) \mid \tau)}_{\log \mathrm{GP}(0, \mathbf{C}(\tau))} P(\tau)$$

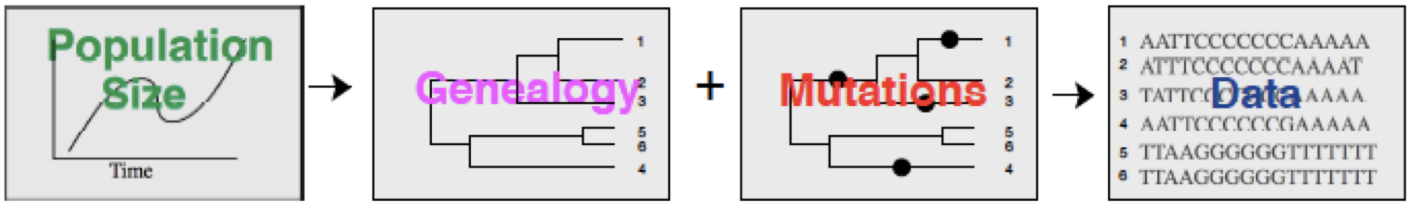
Posterior

Likelihood

Coalescent prior
Birth-death prior

$\log \mathrm{GP}(0, \mathbf{C}(\tau))$
Piece-wise constant (deterministic)
Parametric prior

Bayesian Evolutionary Analysis by Sampling Trees



$N_e(t)$ $\quad$ $\mathbf{G}$ $\quad$ $\mathbf{Q}$ $\quad$ $\mathbf{Y}$

$$P(N_e(t), \mathbf{G}, \mathbf{Q}, \tau \mid \mathbf{Y}) \propto \underbrace{P(\mathbf{Y} \mid \mathbf{G}, \mathbf{Q})}_{\text{Likelihood}} \underbrace{P(\mathbf{G} \mid N_e(t))}_{\text{Coalescent prior}} P(\mathbf{Q}) \underbrace{P(N_e(t) \mid \tau)}_{\log \mathrm{GP}(0, \mathbf{C}(\tau))} P(\tau)$$

Posterior

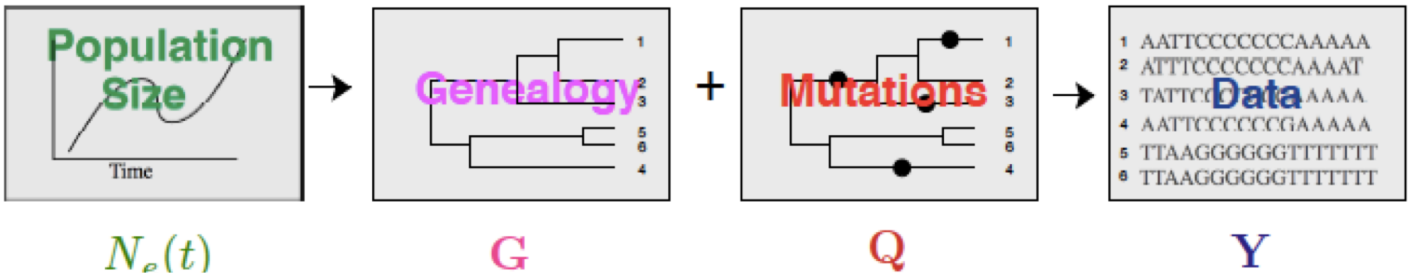
Birth-death prior

Piece-wise constant (deterministic)
Parametric prior

Target of interest: $p(\theta|Y) = \dfrac{p(Y|\theta)p(\theta)}{p(Y)}$

- $p(\theta)$ and $p(Y|\theta)$ – easy
- $p(Y) = \int p(Y|\theta)p(\theta)\mathrm{d}\theta$ – hard

Bayesian Evolutionary Analysis by Sampling Trees



- Goal: $P\left(N_e(t), \mathbf{G}, \mathbf{Q}, \tau \mid \mathbf{Y}\right)$
- The likelihood $P\left(\mathbf{Y} \mid \mathbf{G}, \mathbf{Q}\right)$ is tractable.

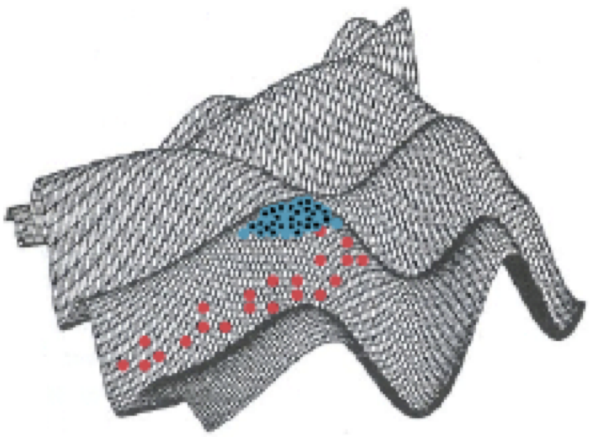
The state space of genealogies $\mathcal{G}$

- $\mathcal{G} = \mathcal{T}_n \times \mathbb{R}_+^{n-1}$
- $|\mathcal{T}_n| = n!(n-1)!/2^{n-1}$
- $|\mathcal{T}_{100}| \approx 10^{284}$

Trouble: $p(Y)$ is not computable – sum over all possible trees

Markov Chain Monte Carlo

- Algorithm generates a Markov chain that visits parameter values (e.g., a specific tree) with frequency equal to their posterior density / probability.
- Markov chain: random walk where the next step only depends on the current parameter state



Marc Suchard – Past SISMID

Metropolis-Hastings Algorithm

- Each step in the Markov chain starts at its current state θ and proposes a new state $\theta^\star$ from an **arbitrary** proposal distribution $q(\cdot|\theta)$ (transition kernel)
- $\theta^\star$ becomes the new state of the chain with probability:

$$
\begin{aligned}
R &= \min\left(1, \frac{p(\theta^\star|Y)}{p(\theta|Y)} \times \frac{q(\theta|\theta^\star)}{q(\theta^\star|\theta)}\right) \\
&= \min\left(1, \frac{p(Y|\theta^\star)p(\theta^\star) \,/\, p(Y)}{p(Y|\theta)p(\theta) \,/\, p(Y)} \times \frac{q(\theta|\theta^\star)}{q(\theta^\star|\theta)}\right) \\
&= \min\left(1, \frac{p(Y|\theta^\star)p(\theta^\star)}{p(Y|\theta)p(\theta)} \times \frac{q(\theta|\theta^\star)}{q(\theta^\star|\theta)}\right)
\end{aligned}
$$

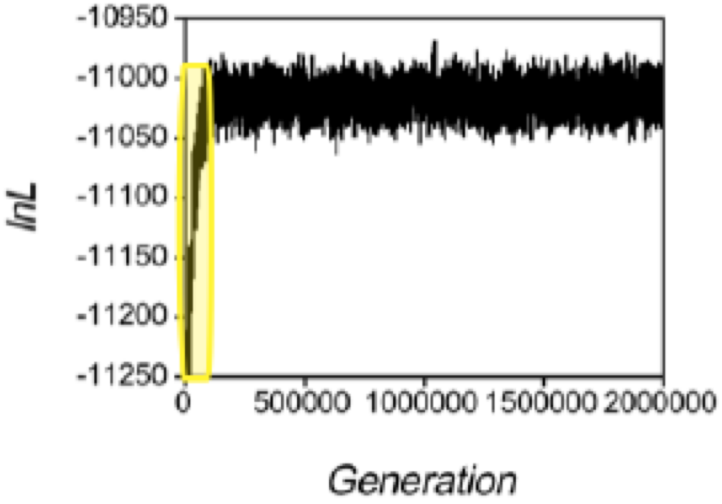
- Otherwise, θ remains the state of the chain

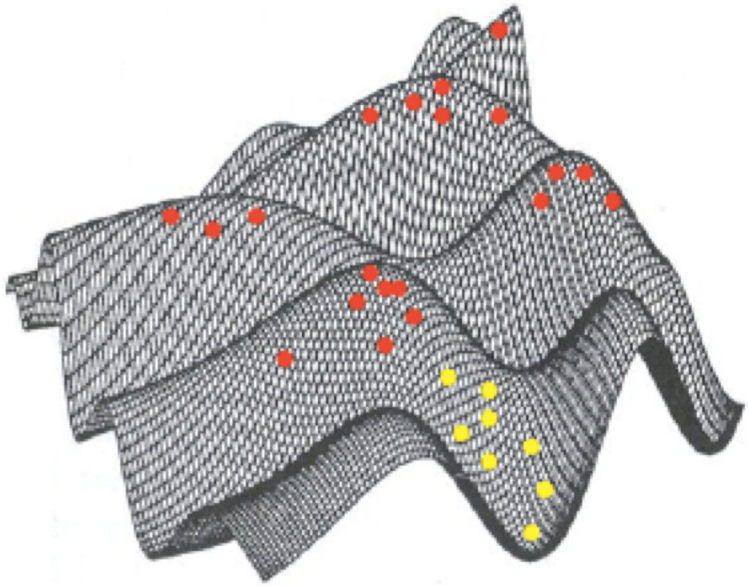
Marc Suchard – Past SISMID

Metropolis-Hastings Algorithm

We repeat the process of proposing a new state, calculating the acceptance probability and either accepting or rejecting the proposed move millions of times

Although correlated, the Markov chain samples are valid draws from the posterior; however ...

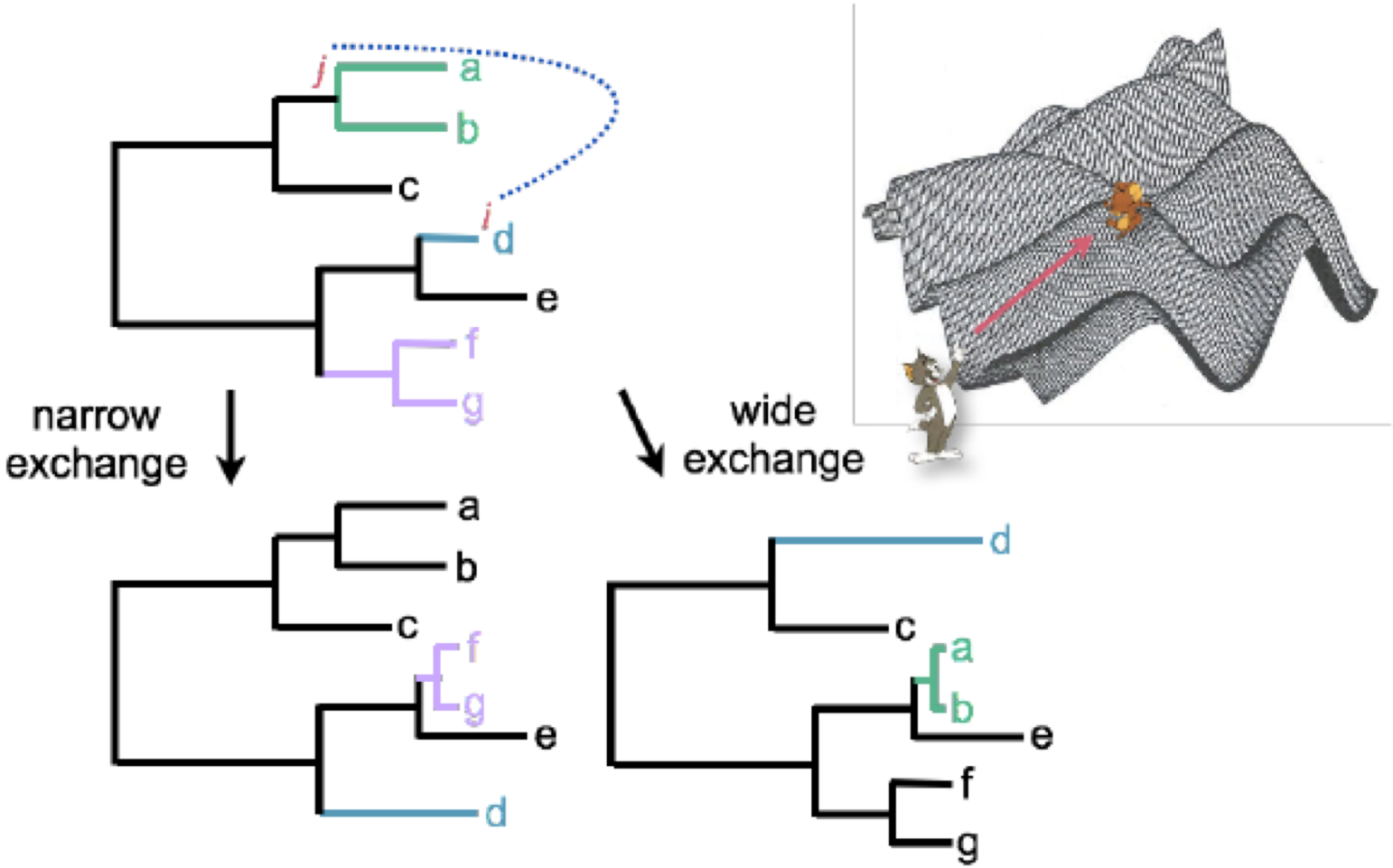




Initial sampling (burn-in) is often discarded due to correlation with chain's starting point ($\neq$ posterior)

Marc Suchard – Past SISMID

Transition kernels

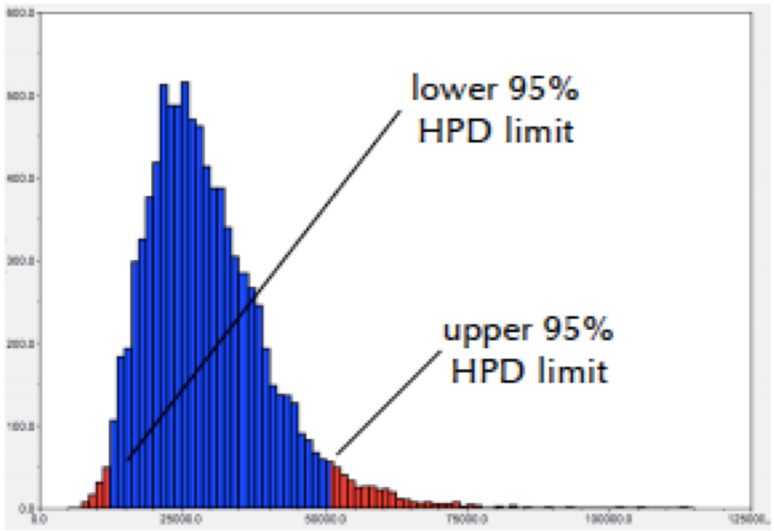


Marc Suchard – Past SISMID

Posterior summaries

For continuous θ, consider:

- posterior mean or median ≈ MCMC sample average or median
- quantitative measures of uncertainty, e.g. high posterior density interval



For trees, consider:

- scientifically interesting posterior probability statement, e.g. the probability of monophyly ≈ MCMC sample proportion under which hypothesis is true

Book references