

Evolutionary Dynamics and Molecular Epidemiology of Viruses

Instructors: Julia A. Palacios and Nicola Müller

TAs: Nídia Trovão and Joëlle Barido-Sottani

Program

Monday, July 27

- Introduction
- Tutorial: Introduction to BEAST 2
- Tracking evolution through time
- Tutorial: Convergence and Troubleshooting

Tuesday, July 28

- Tree priors.
- Tutorial: Estimation of effective population size.
- Structure populations
- Tutorial: Structured coalescent and BD models
- Tree Distances, Convergence and Visualization

Wednesday, July 29

- Recombination and Reassortment
- Tutorial: Coalescent with reassortment

Logistics and Questions

- 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.

Introduction to evolutionary dynamics and molecular epidemiology of viruses

The goal is to understand:

- Patterns of transmission and spread (effective population size)
- Mutation rate
- Comparison to other pathogens
- Sources of molecular variation (mutation, selection, recombination)
- Surveillance

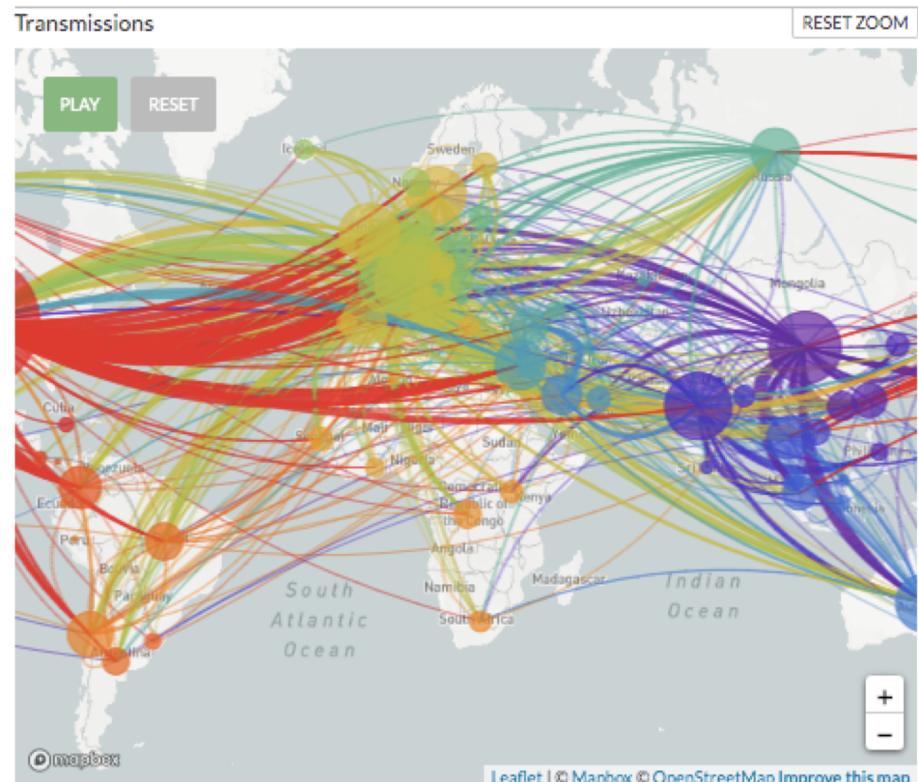
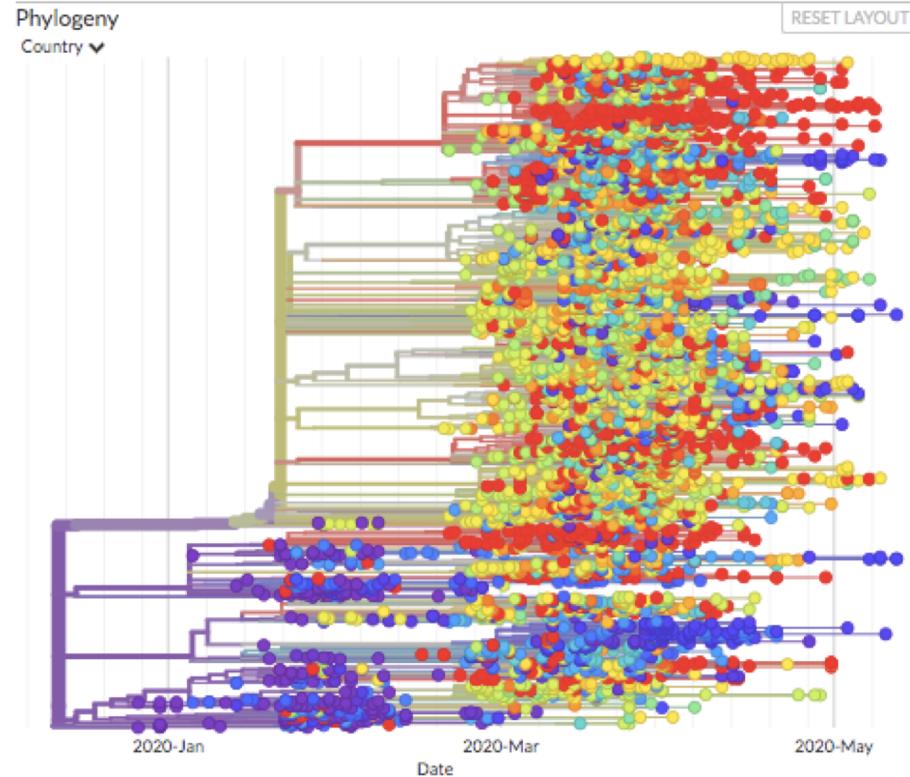
Molecular epidemiology and **phylodynamics** of infectious diseases aim to study infectious disease behavior through a combination of evolutionary, epidemiological and immunological processes from molecular variation [HG09].

Global spread of SARS-CoV-2

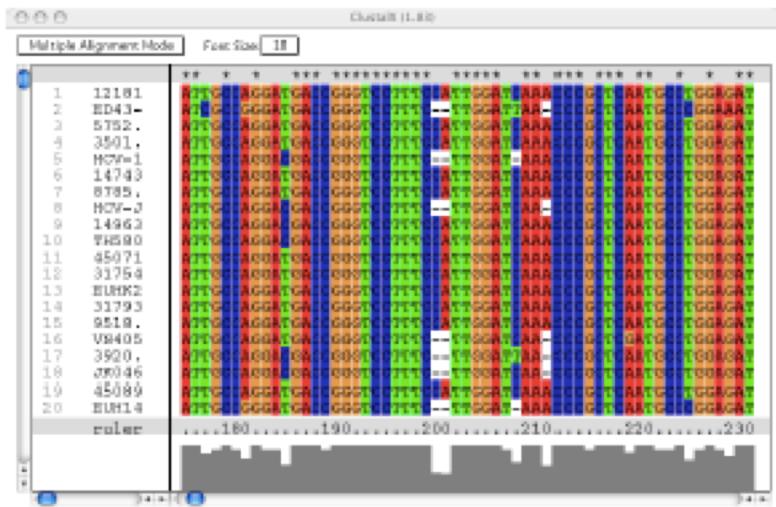
Genomic epidemiology of novel coronavirus - Global subsampling

Maintained by the Nextstrain team. Enabled by data from [GISAID](#)

Showing 4256 of 4256 genomes sampled between Dec 2019 and May 2020.



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.

Genealogies and Phylogenies

PNAS

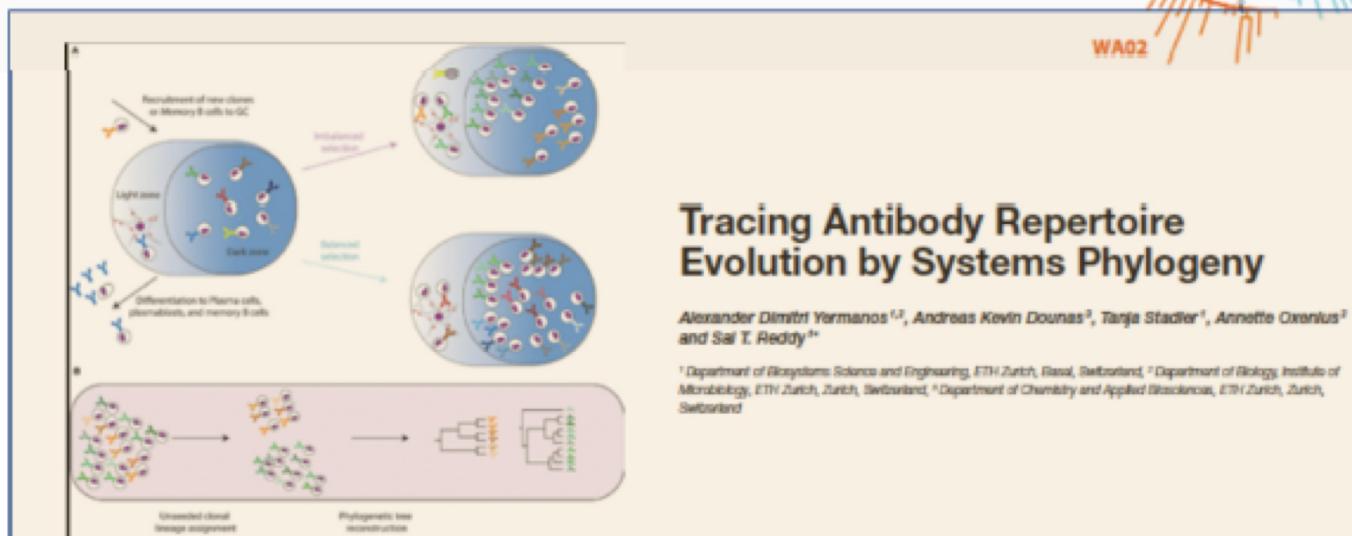
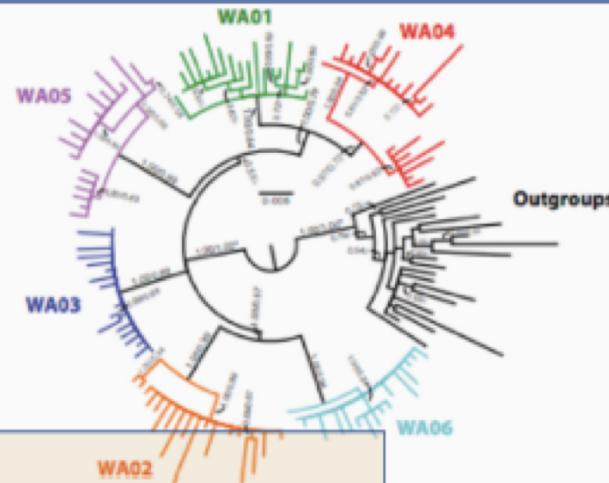
Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences

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

^aHuman Genome Sequencing Center, ^bDepartment of Molecular and Human Genetics, and ^cCell and Molecular Biology Program, Baylor College of Medicine, Houston, TX 77030; and ^dSection of Integrative Biology and Center for Computational Biology and Bioinformatics, University of Texas, Austin, TX 78712

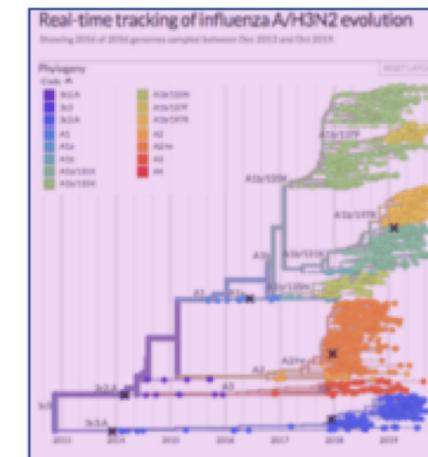
This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2008.

Contributed by David M. Hillis, October 26, 2010 (sent for review September 22, 2010)



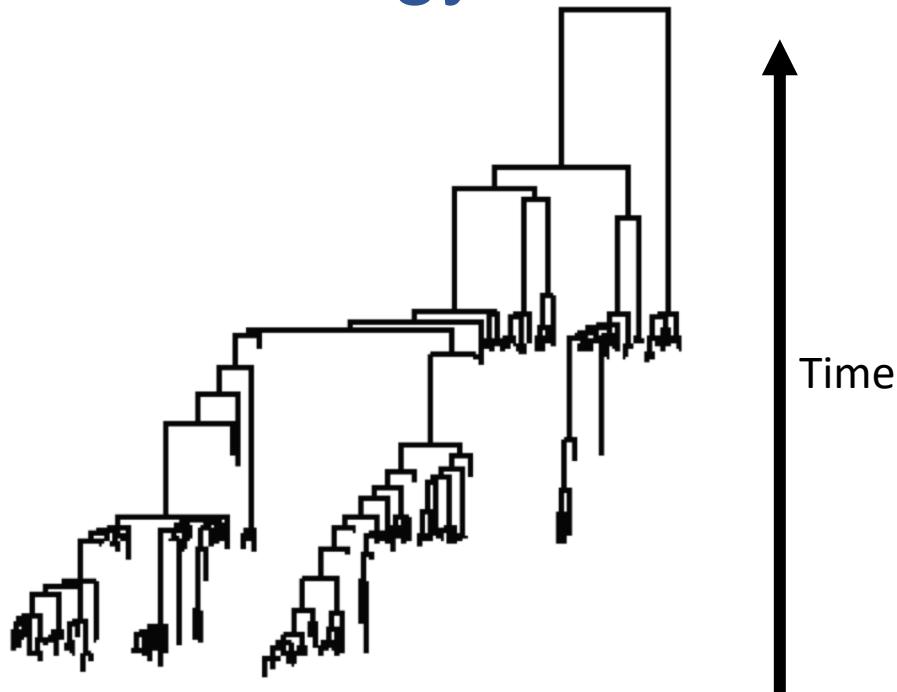
Alexander Dimitri Yermanov^{1,2}, Andreas Kevin Dounas³, Tanja Stadler¹, Annette Oxentus², and Sai T. Reddy^{1*}

¹Department of Bioengineering Sciences and Engineering, ETH Zurich, Riesal, Switzerland; ²Department of Biology, Institute of Microbiology, ETH Zurich, Zurich, Switzerland; ³Department of Chemistry and Applied Biosciences, ETH Zurich, Zurich, Switzerland



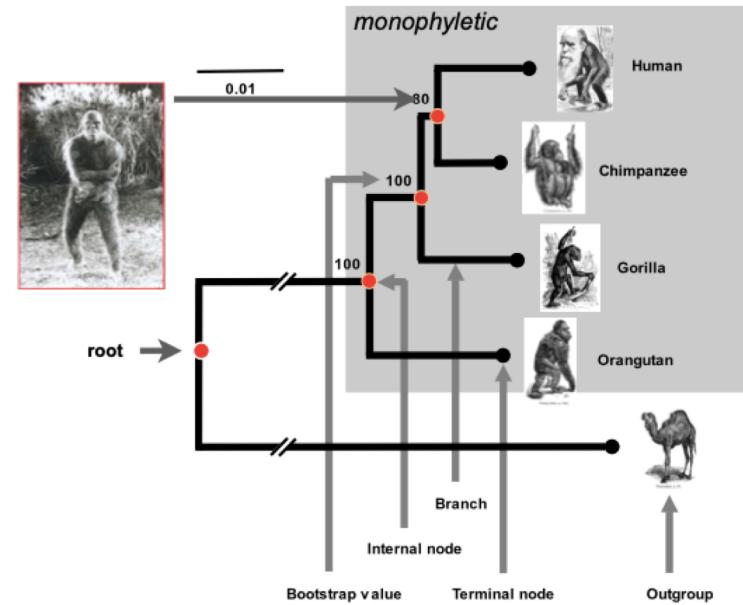
What is a genealogy?

Genealogy



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

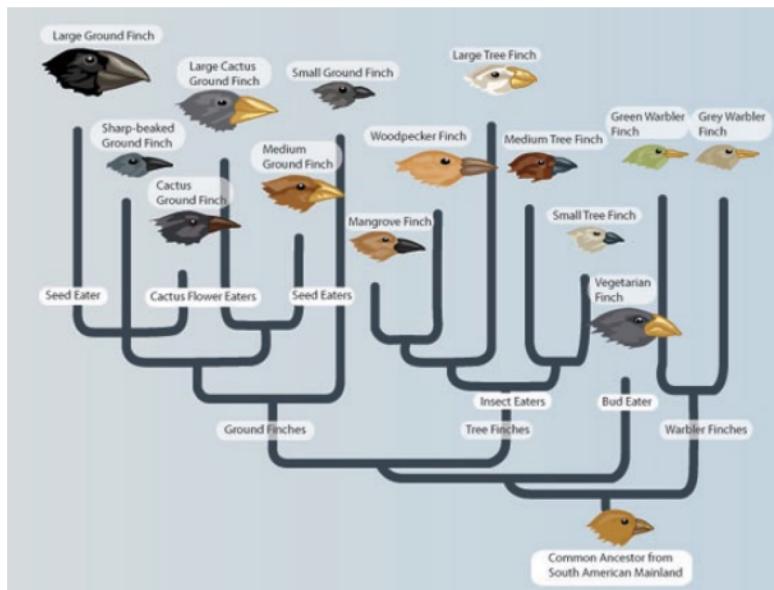
Phylogeny



- 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

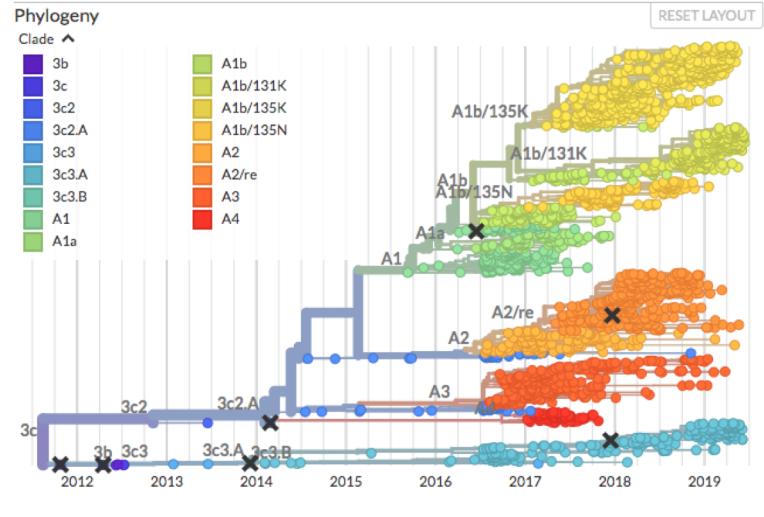


- **Phylodynamics**

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

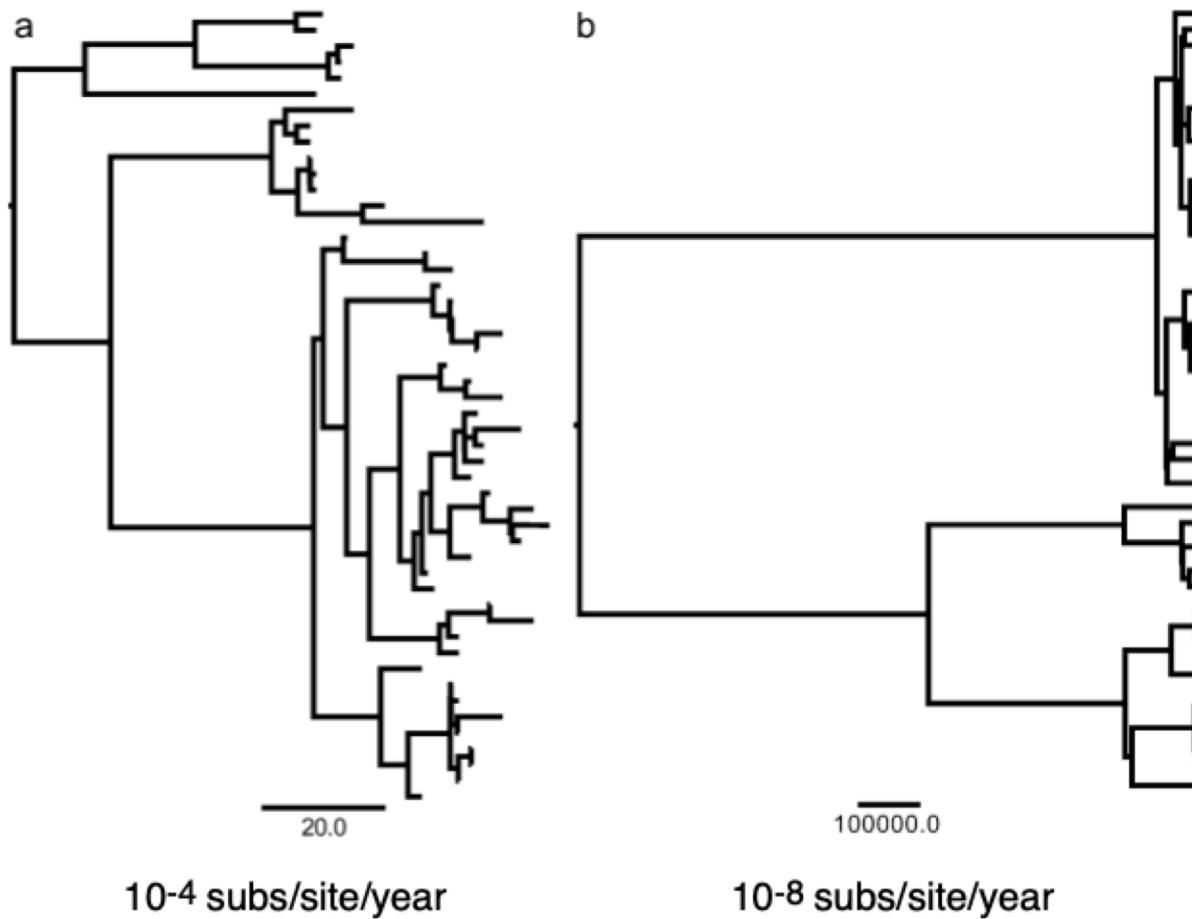
Real-time tracking of influenza A/H3N2 evolution

Showing 2169 of 2169 genomes sampled between Oct 2011 and Jun 2019 and comprising 17 clade member

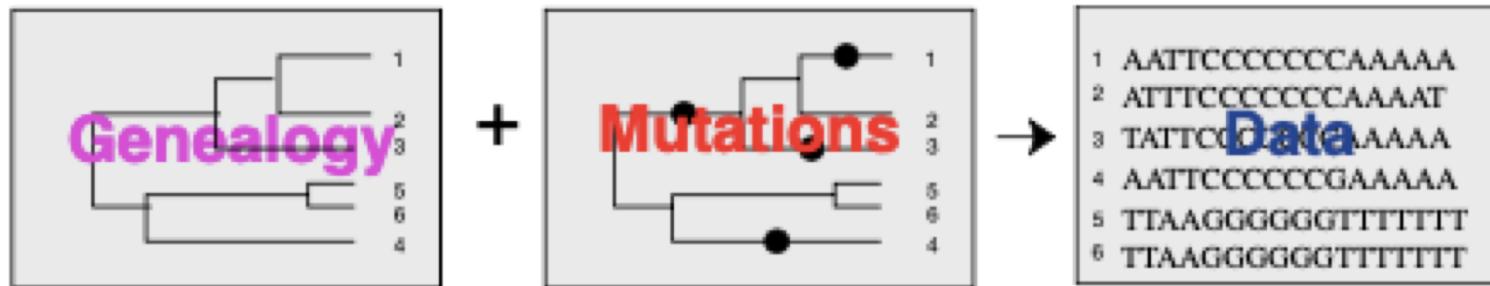


Phylogenetics and population genetics

- For rapidly evolving organisms
- For slowly evolving organisms

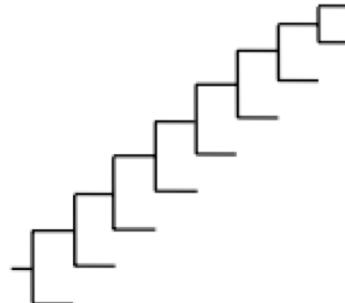
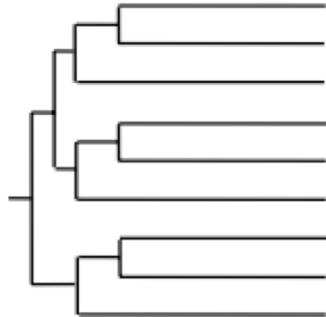
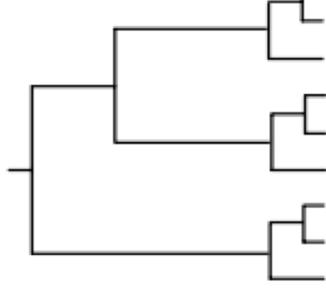
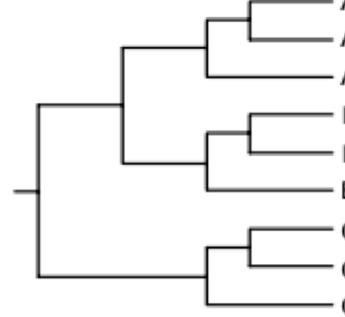
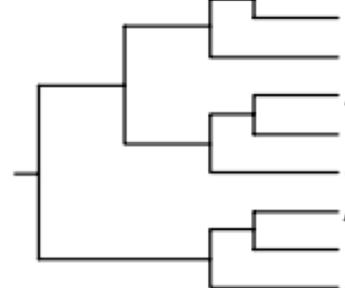


Statistical Phylogenetics seeks to infer genealogies from molecular data



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

Phylogenetic Patterns

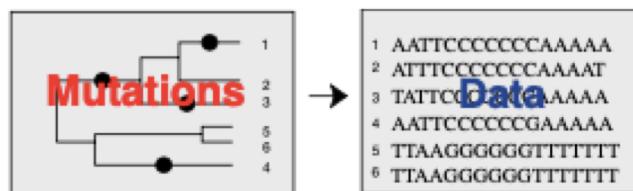
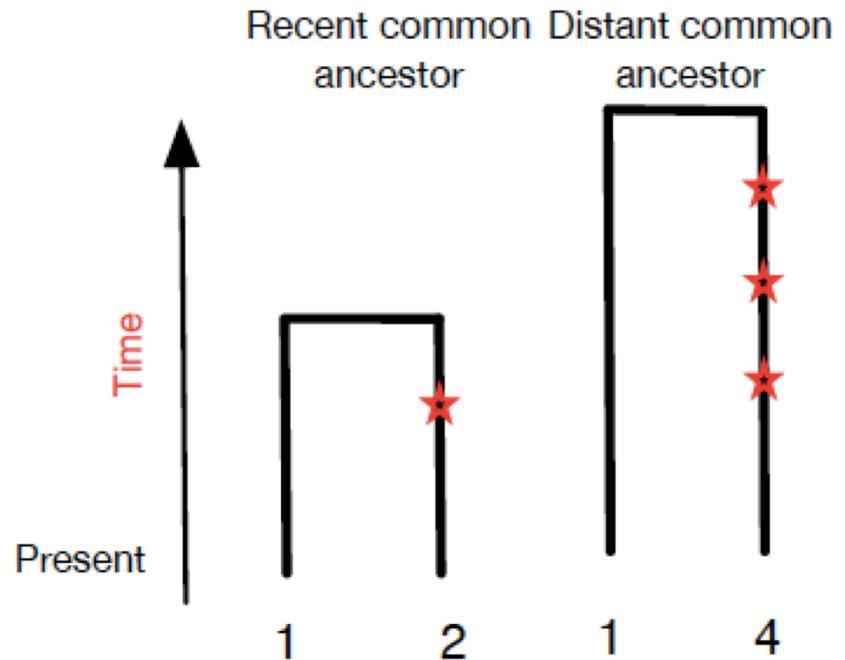
	Continual Immune Selection	Weak/No Immune Selection	
Idealised Phylogeny Shapes		Population dynamics	Spatial dynamics
		<i>Population growth</i>  <i>Population decline</i> 	<i>Strong spatial structure</i>  <i>Weak spatial structure</i> 
Examples	Human influenza A within-host HIV	among-host HIV among-host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

Phylogenetic Reconstruction

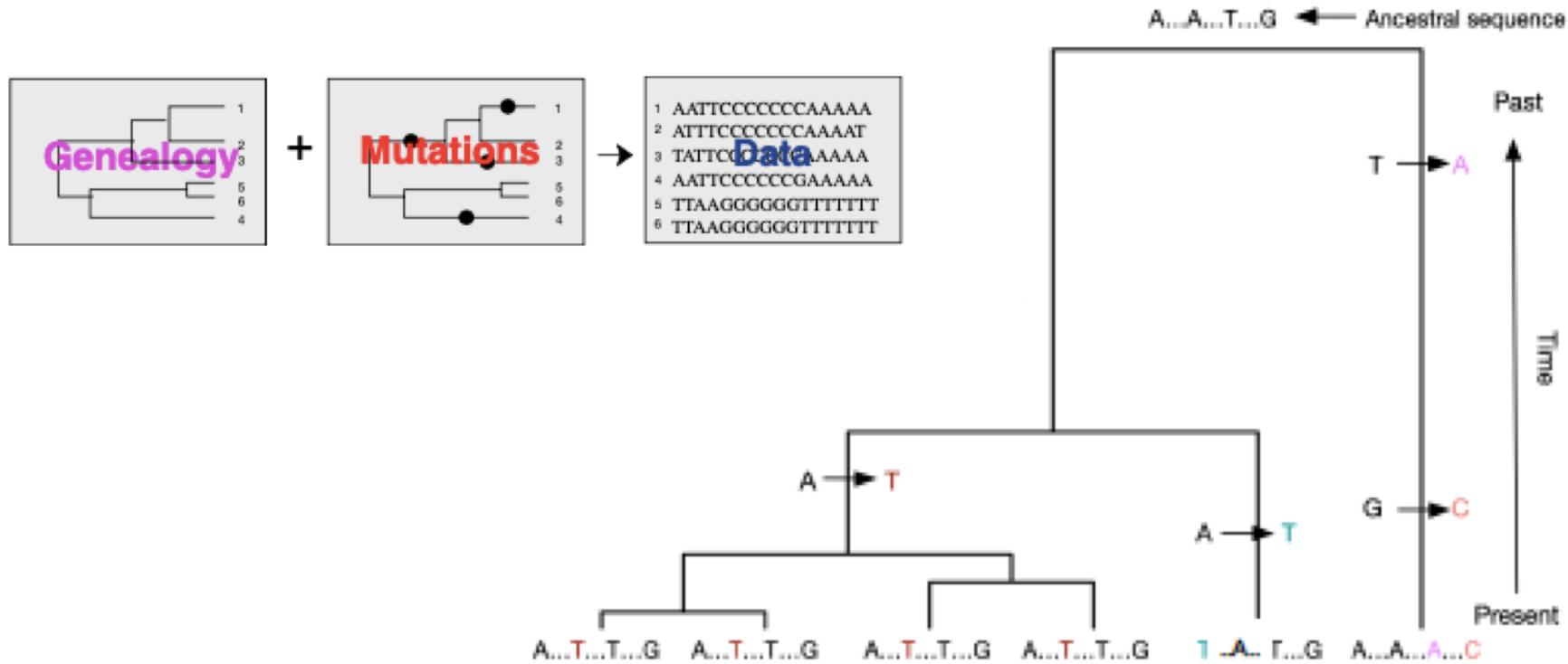
- Clustering Approaches
 - These begin with a genetic distance between each pair of sequences. A ‘clustering algorithm’ then transforms the genetic distances into a tree.
 - UPGMA, Neighbour-joining
 - + Simple, faster
 - No uncertainty assessment or statistical guarantees
- Optimality and Statistical methods
 - These define a score for each possible tree
 - Parsimony, Maximum Likelihood, Maximum a posteriori (**Bayesian**)
 - More complex, slower.
 - Statistical guarantees and uncertainty quantification.

Observed differences inform about ancestry

1	ATTCAACCGG
2	ATTCAACCTG
3	TTC CAACCTG
4	ATTCA CGCTG

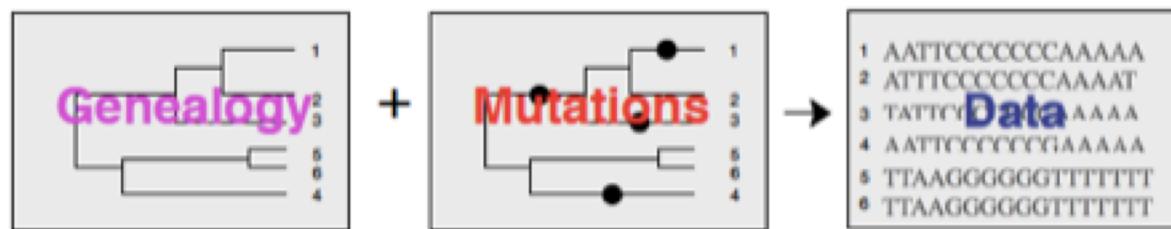


A process of substitutions superimposed on the genealogy generates observed sequences at the tips of the genealogy



Statistical Phylogenetics

Goal: Estimate genealogy/phylogeny



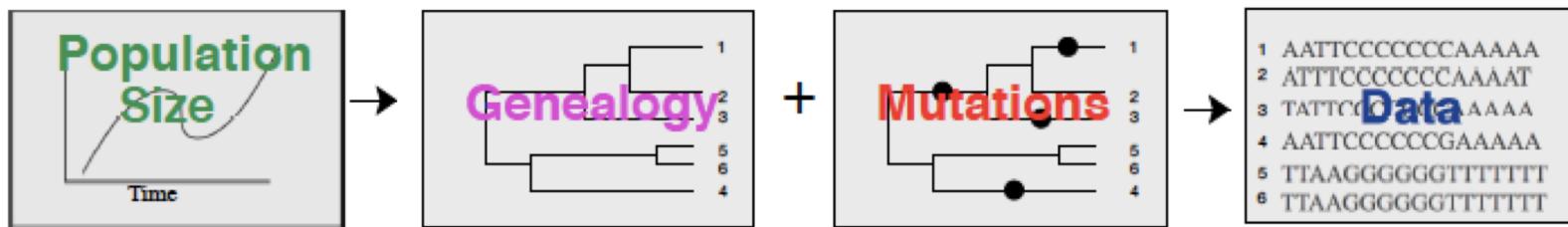
Phyldynamics

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



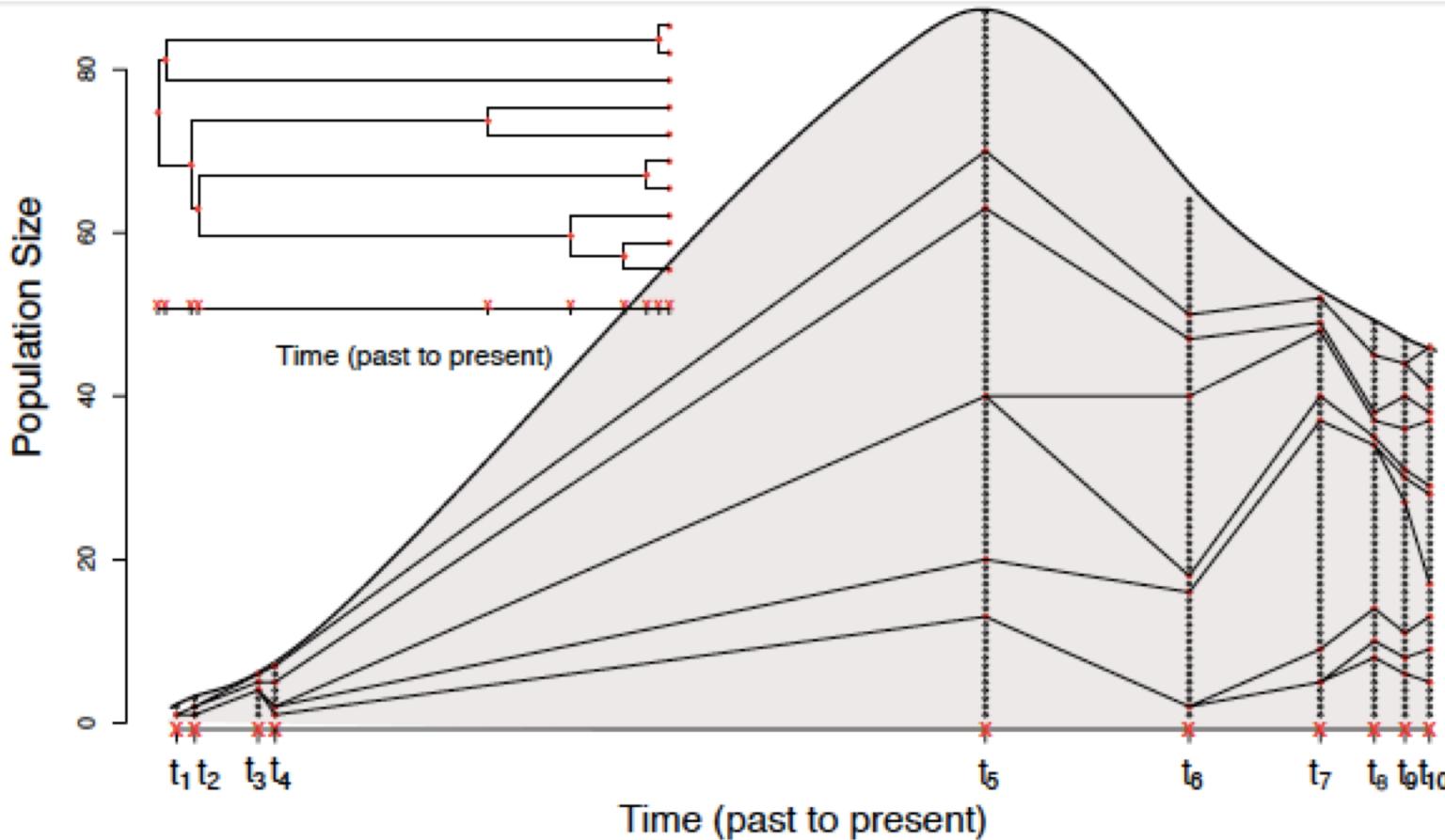
Goal: Estimation of Effective Population Size

Coalescent-based model



- ▶ Ancestral process: **coalescent process** of genealogies.
- ▶ Mutation process: **CTMC** along the branches of the genealogy ($HKY + CP_{112} + \Gamma_{112}$).
- ▶ Population process: **Effective population size** trajectory over time.

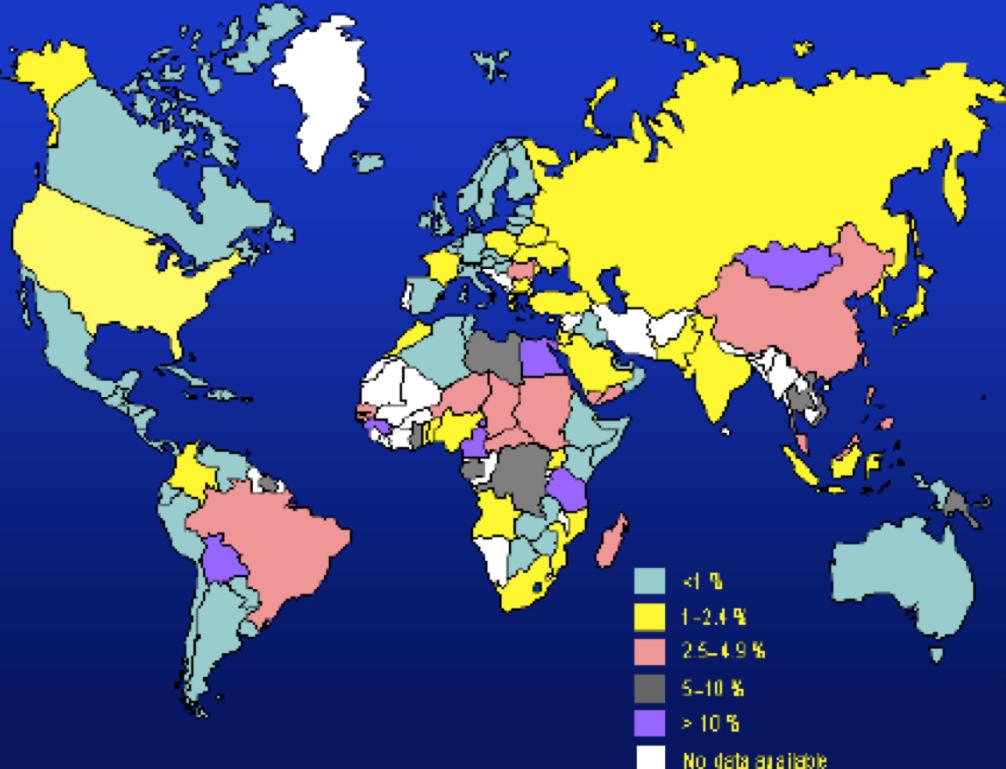
From genealogy to population size



Long waiting times to find a common ancestor indicates large population sizes

Example 1: Hepatitis C in Egypt

HCV Has Broad Global Prevalence



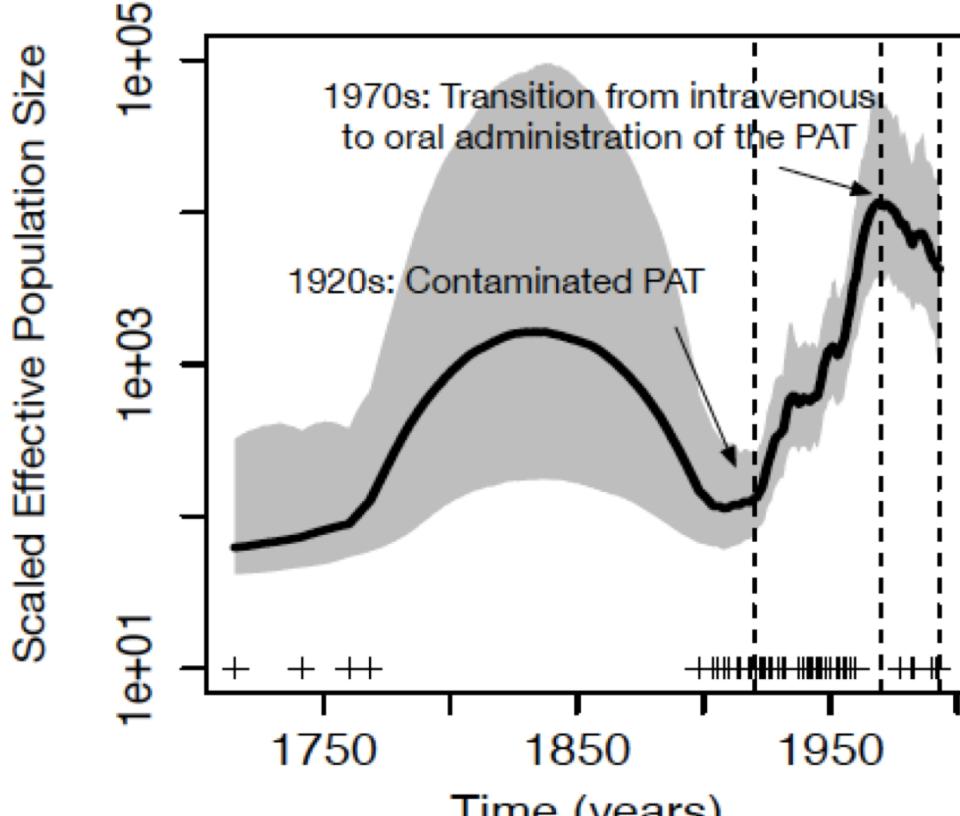
Prevalence of HCV - WHO 1999

- 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

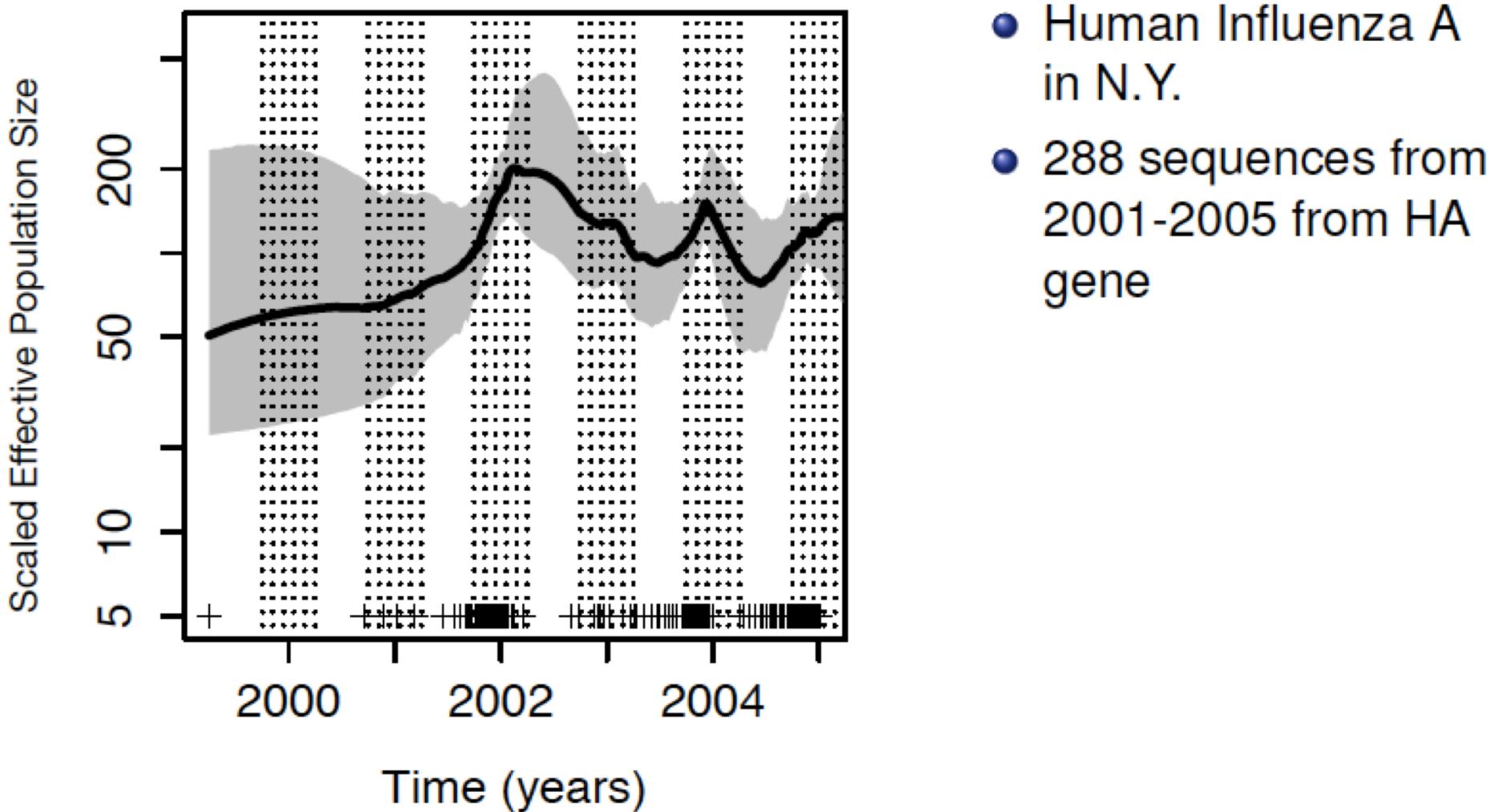
Example 1: Hepatitis C in Egypt



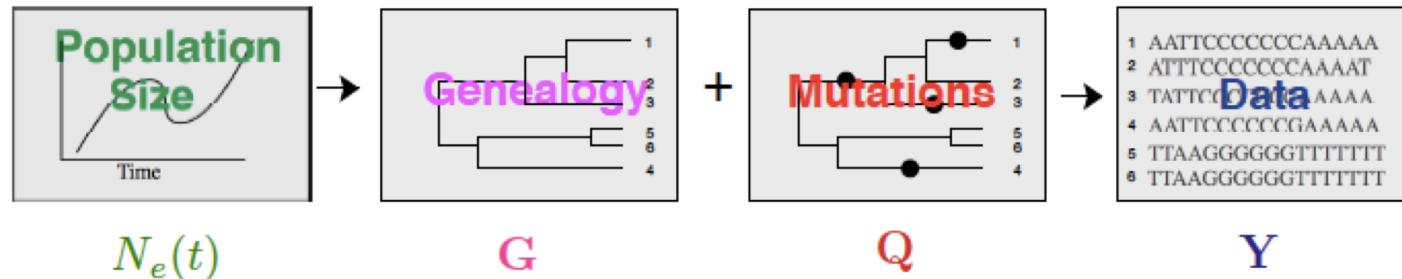
[Palacios and Minin, Biometrics 2013]

- 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



Bayesian Evolutionary Analysis by Sampling Trees



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

Birth-death prior
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.

$$\pi(\theta) \quad (1)$$

This express your belief about θ before you have seen the data.

- The sampling distribution (or likelihood) has a known functional form: $L(X_1, \dots, X_n | \theta)$.
- Applying Bayes' rule, we get the following posterior distribution

$$P(\theta | X_1, \dots, X_n) = \frac{L(X_1, \dots, X_n | \theta)\pi(\theta)}{\int_{\theta \in \Theta} L(X_1, \dots, X_n | \theta)\pi(\theta)d\theta} \quad (2)$$

Bayesian Inference

$$\pi(\theta) \quad (3)$$

$$L(X_1, \dots, X_n \mid \theta) \quad (4)$$

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"

$$P(\theta \mid X_1, \dots, X_n) \quad (5)$$

Example: Poisson-Gamma

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

$$P(\theta \mid x_1, \dots, x_n) = \frac{P(x_1, \dots, x_n \mid \theta) P(\theta)}{P(x_1, \dots, x_n)}$$

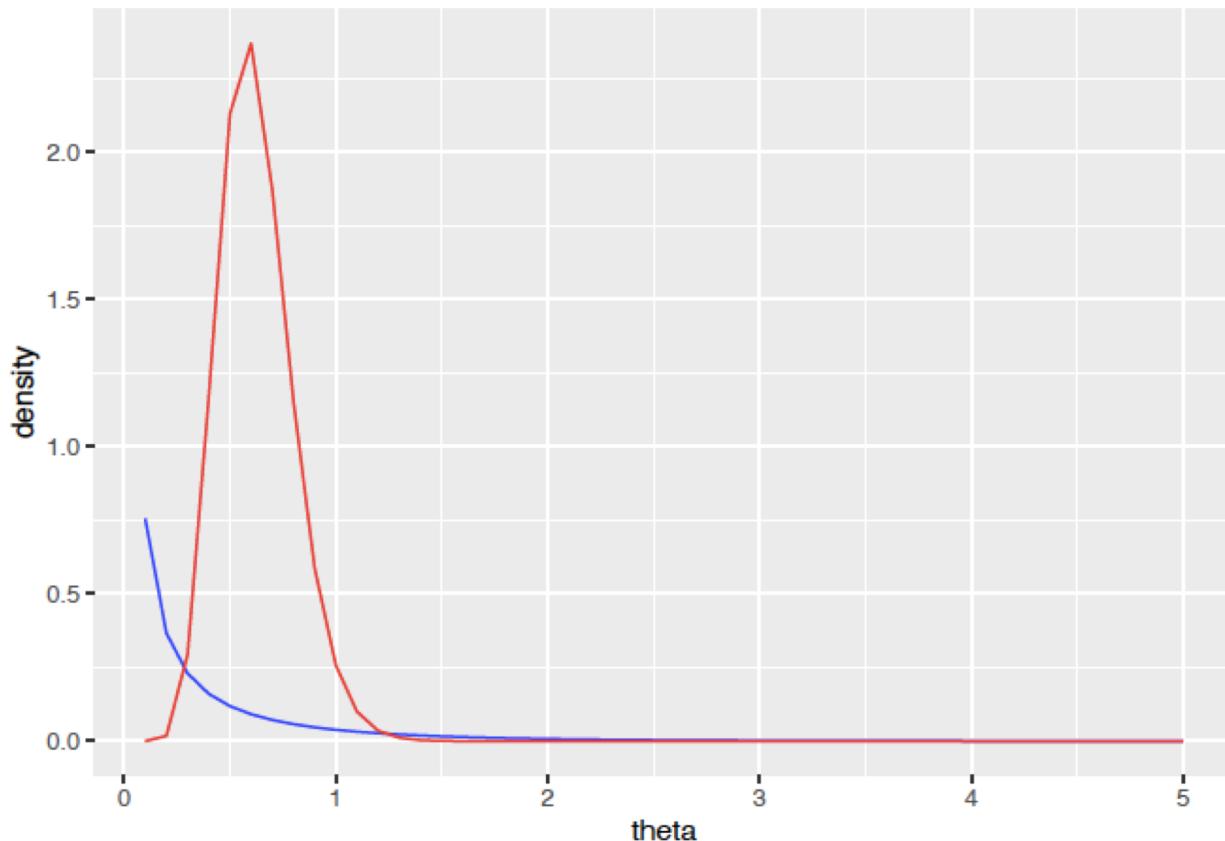
$$P(\theta \mid x_1, \dots, x_n) = \frac{\theta^{\sum_i^n x_i} e^{-\theta} (\prod_{i=1}^n x_i!)^{-1} \theta^{\alpha-1} e^{-\theta/\beta} (\Gamma(\alpha)\beta^\alpha)^{-1}}{P(x_1, \dots, x_n)}$$

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

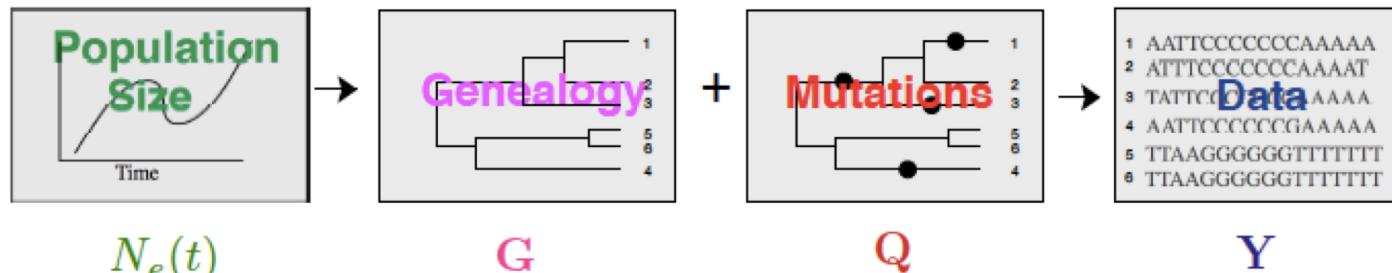
Example: Poisson-Gamma

```
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



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

Piece-wise constant (deterministic)
Parametric prior

A process of substitutions is superimposed on the genealogy

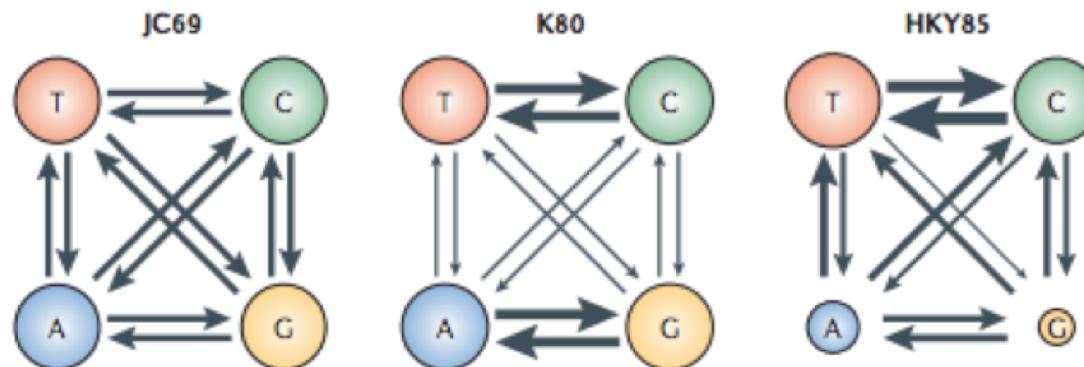
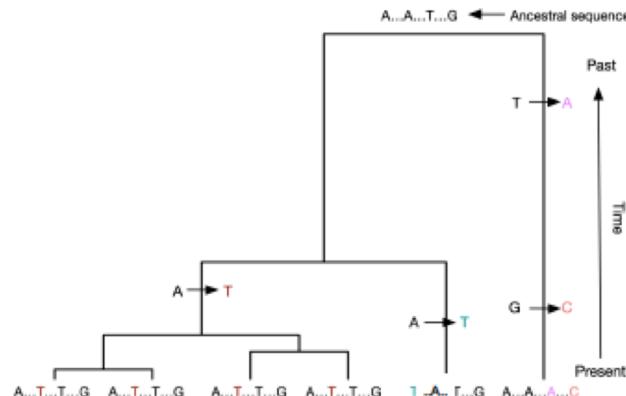
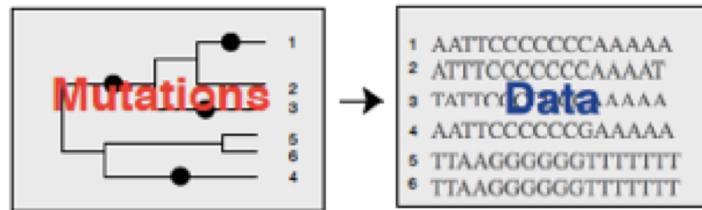
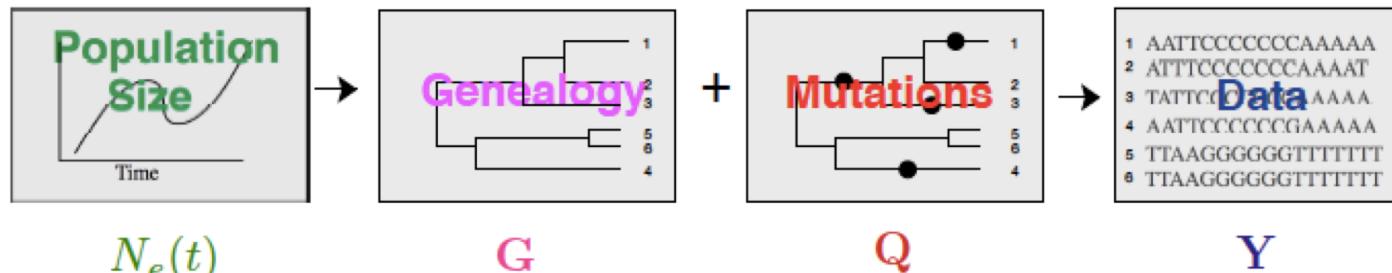


Figure 1 | Markov models of nucleotide substitution. The thickness of the arrows indicates the substitution rates of the four nucleotides (T, C, A and G), and the sizes of the circles represent the nucleotide frequencies when the substitution process is in equilibrium. Note that both JC69 and K80 predict equal proportions of the four nucleotides.

Transversions < transitions

Bayesian Evolutionary Analysis by Sampling Trees



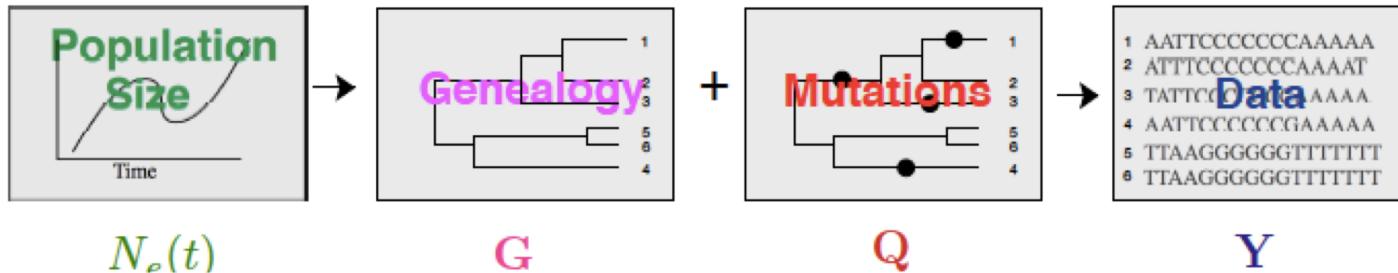
$$P(N_e(t), \mathbf{G}, \mathbf{Q}, \tau | \mathbf{Y}) \propto \underbrace{P(\mathbf{Y} | \mathbf{G}, \mathbf{Q})}_{\text{Posterior}} \underbrace{P(\mathbf{G} | N_e(t))}_{\text{Likelihood}} \underbrace{P(\mathbf{Q})}_{\text{Coalescent prior}} \underbrace{P(N_e(t) | \tau)}_{\log GP(0, C(\tau))} P(\tau)$$

Birth-death prior Piece-wise constant (deterministic)
Parametric prior

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

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

Bayesian Evolutionary Analysis by Sampling Trees



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

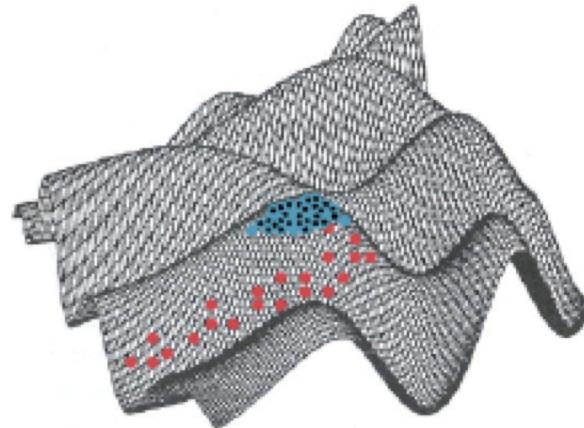
The state space of genealogies \mathcal{G}

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

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



Metropolis-Hastings Algorithm

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

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

- Otherwise, θ remains the state of the chain

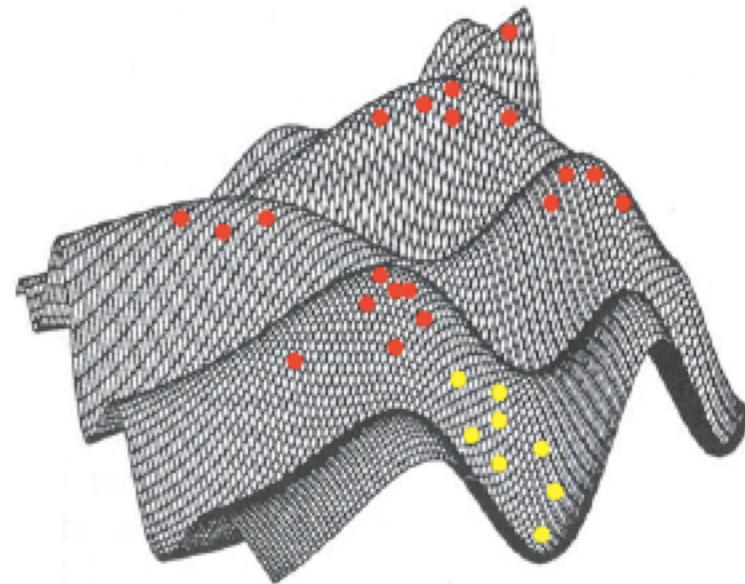
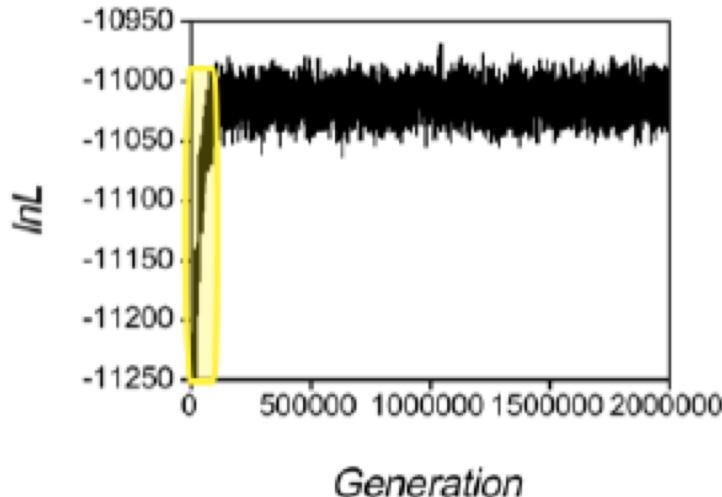
Marc Suchard – Past SISMID

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 . . .

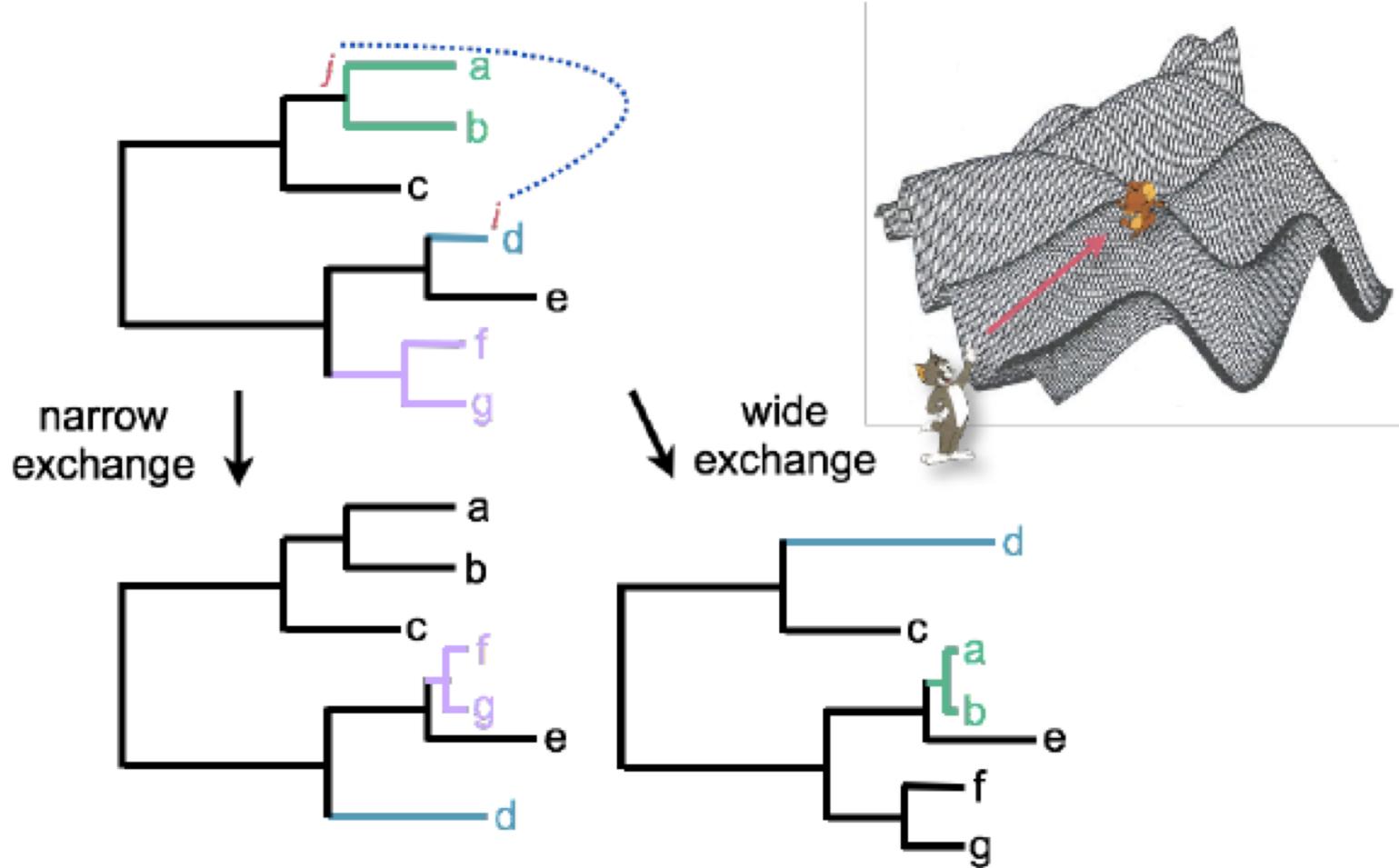


Initial sampling (burn-in) is often discarded due to correlation with chain's starting point (\neq posterior)

Marc Suchard – Past SISMID

Julia Palacios

Transition kernels



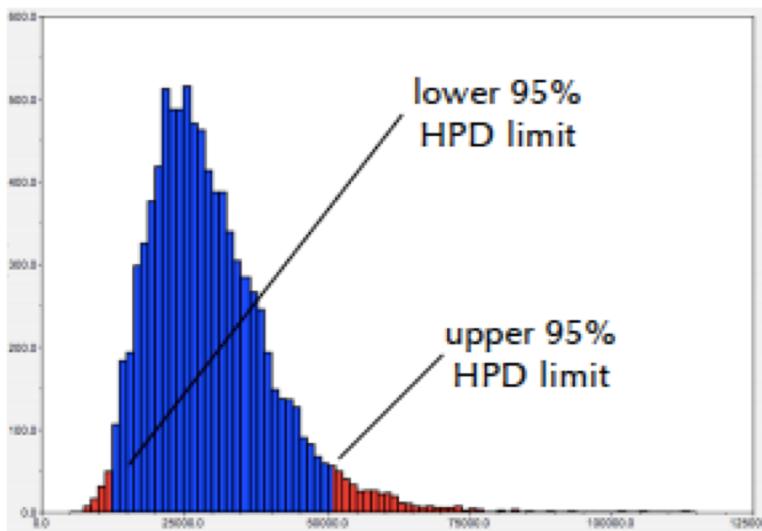
Marc Suchard – Past SISMID

Julia Palacios

Posterior summaries

For continuous θ , consider:

- posterior mean or median \approx 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 \approx MCMC sample proportion under which hypothesis is true



Book references

