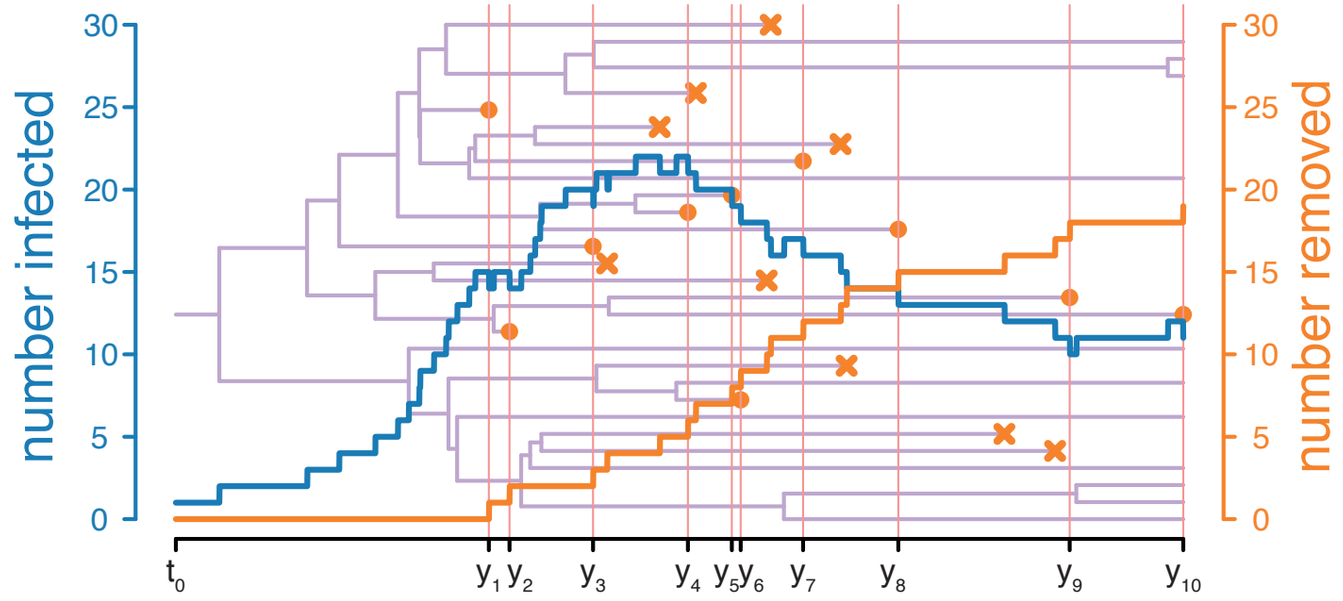
Inferring epidemiological parameters from phylogenies with BEAST2



Swiss Institute of Bioinformatics

Veronika Bošková, Louis du Plessis, Joëlle Barido-Sottani
Computational Evolution (www.bsse.ch/cevo)
Department of Biosystems Science and Engineering

CEVO ETH zürich

Inferring epidemiological parameters from phylogenies with BEAST2

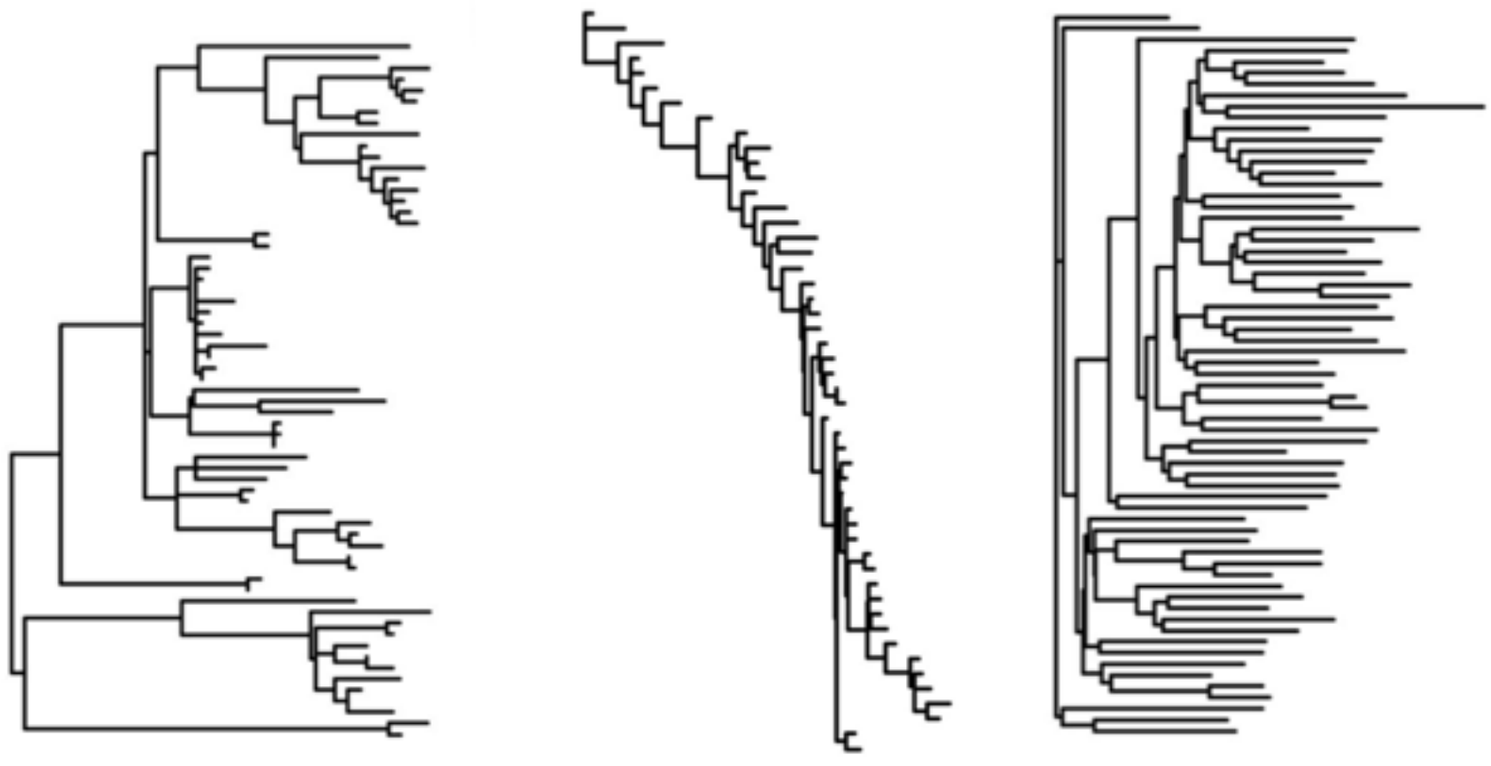
1. Introduction to Phylodynamics
2. Parameter inference
3. Likelihood calculation
4. BEAST2 Howto
5. Tutorial

Veronika Bošková, Louis du Plessis, Joëlle Barido-Sottani
Computational Evolution (www.bsse.ch/cevo)
Department of Biosystems Science and Engineering

Swiss Institute of Bioinformatics

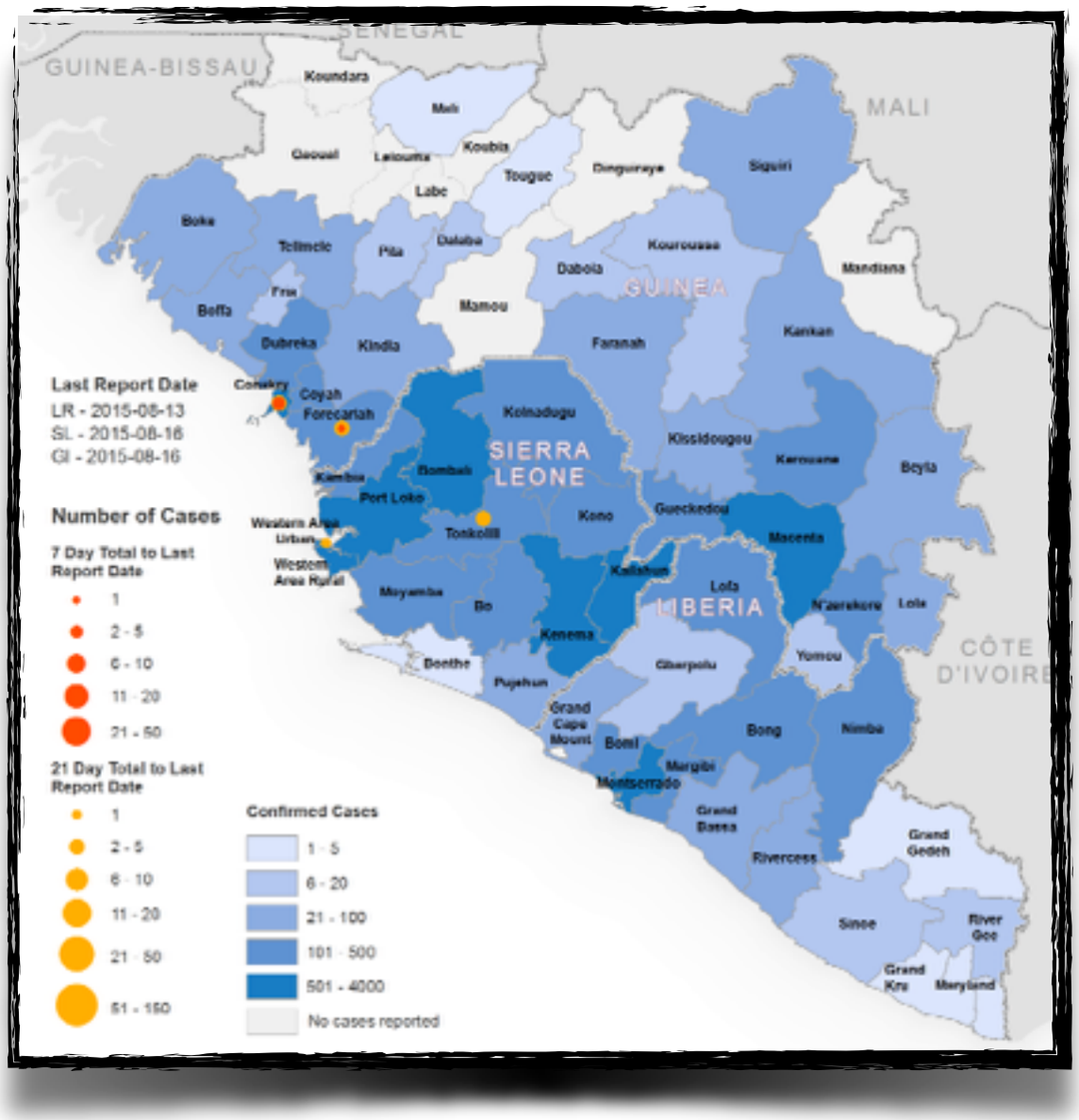
ETH zürich

Section 1: Introduction to Phylodynamics

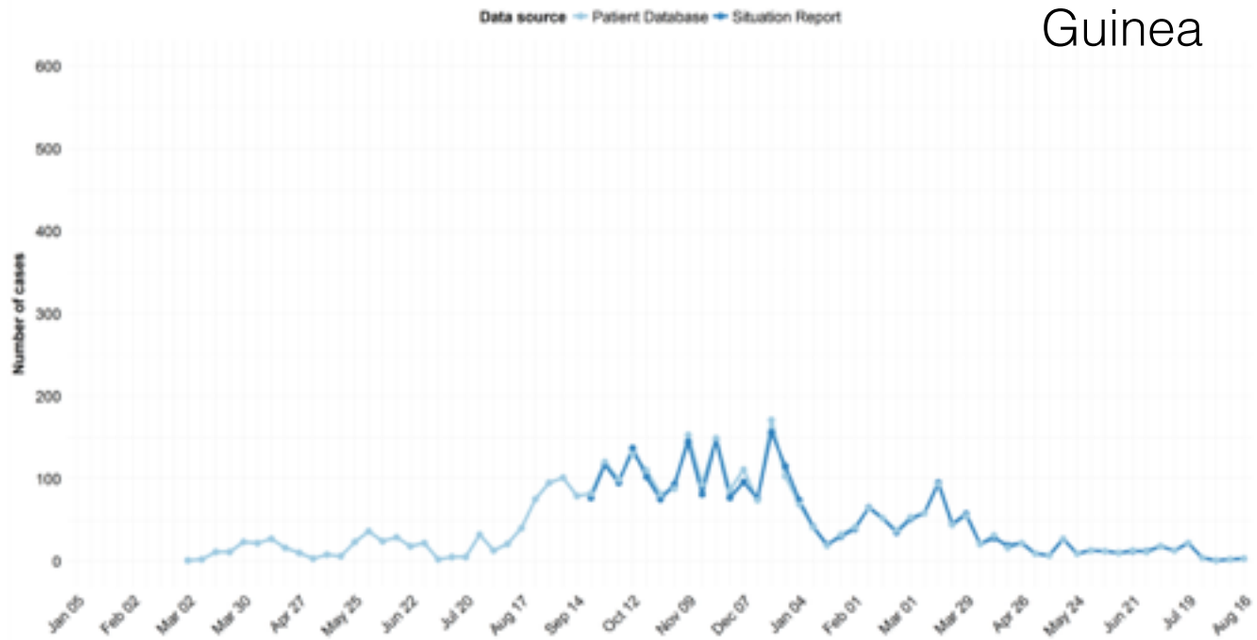


Louis du Plessis

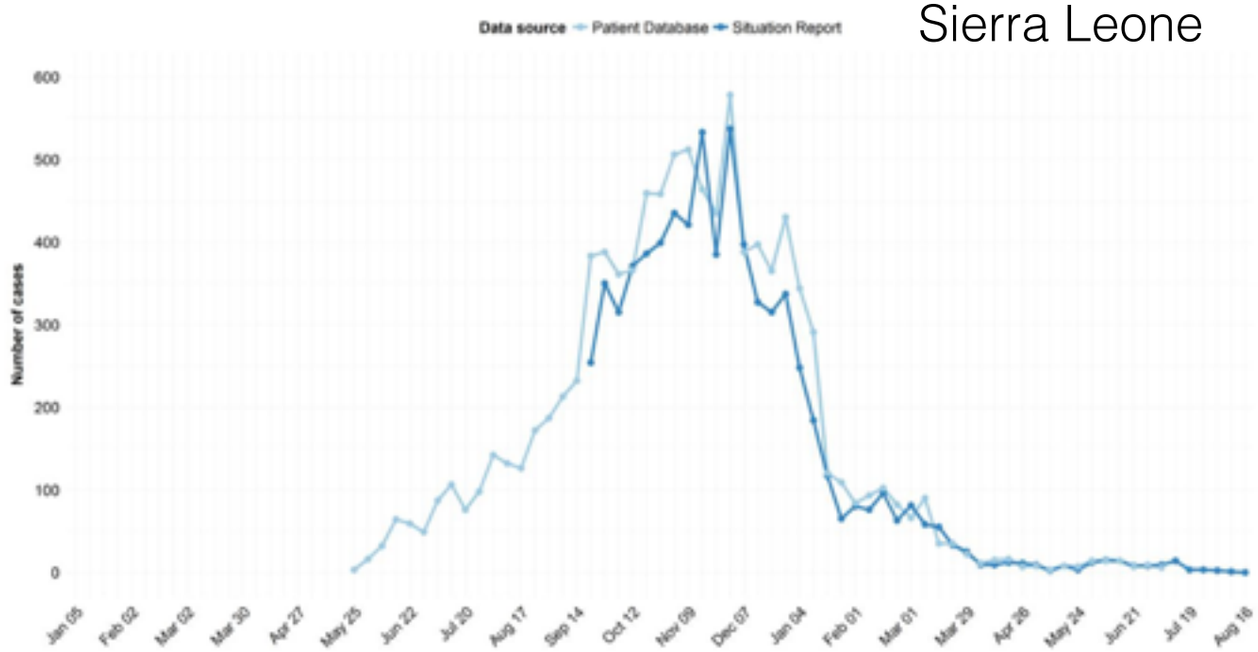
How can we learn about epidemic spread?



Guinea

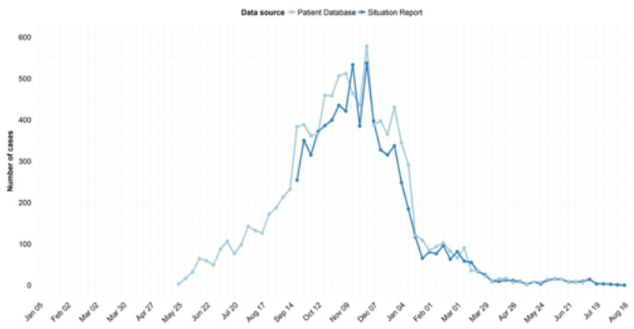


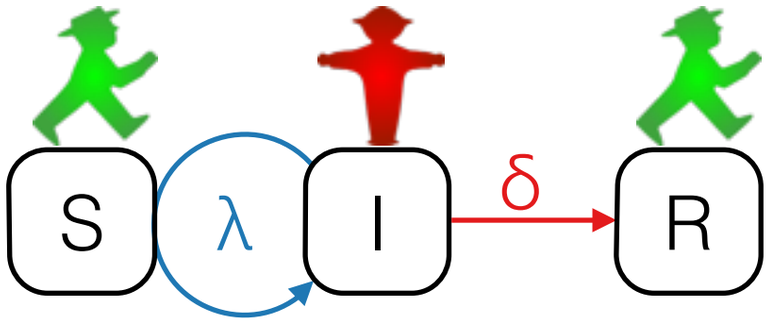
Sierra Leone



http://apps.who.int/ebola

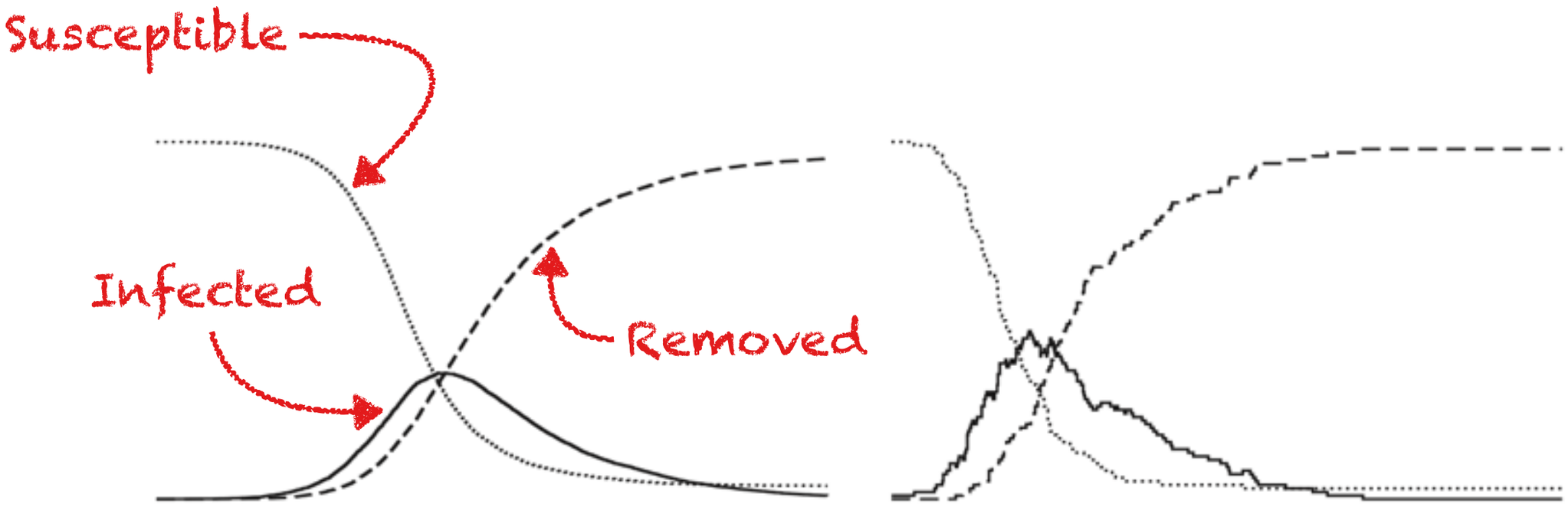
Compartmental models for epidemics





$$dS/dt = -\lambda IS/N$$
$$dI/dt = \lambda IS/N - \delta I$$
$$dR/dt = \delta I$$

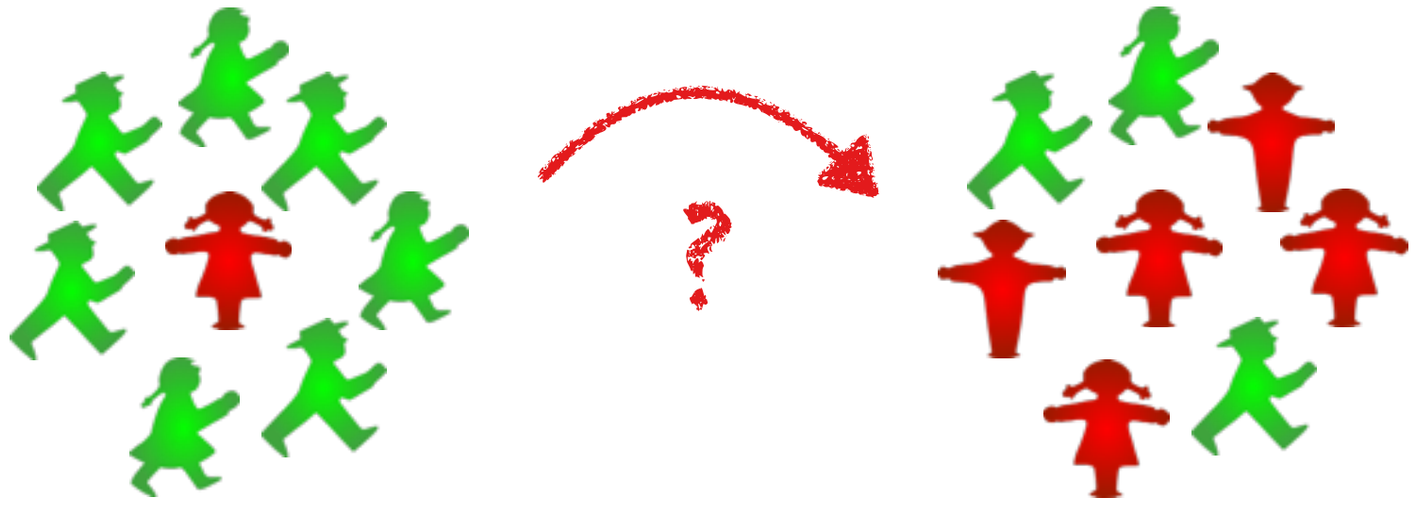
- λ — infection rate
- δ — becoming-noninfectious rate



Deterministic

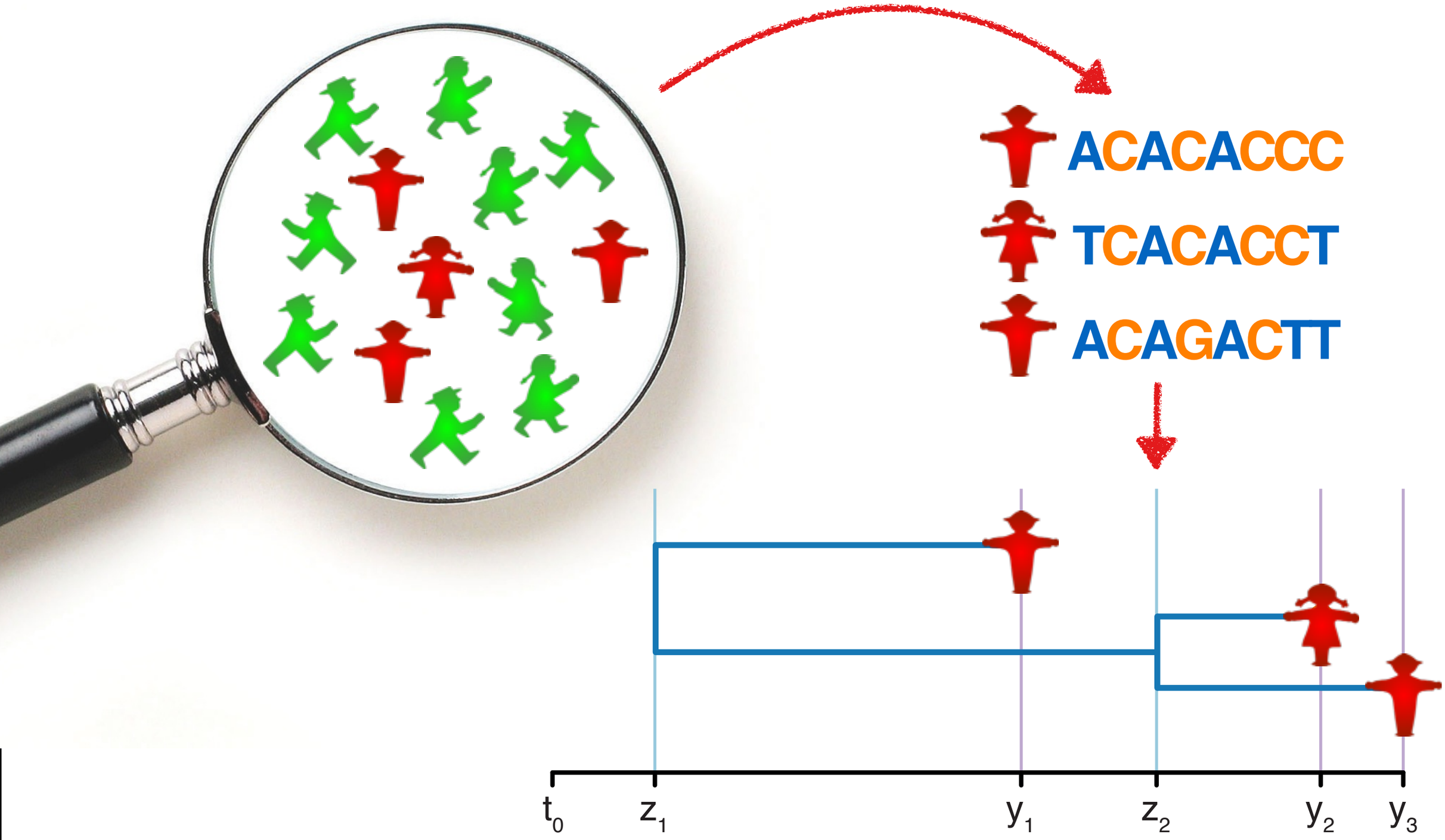
Stochastic

What are the important questions?



- How fast is the epidemic spreading?
  (Basic reproduction number — R_0)
- When/where did the epidemic originate?
- Are public health interventions effective?
  (Effective reproduction number — R_e)
- Does population structure play a role?
- Did we sample everyone who is infected?
- …?

Can phylogenetics help?



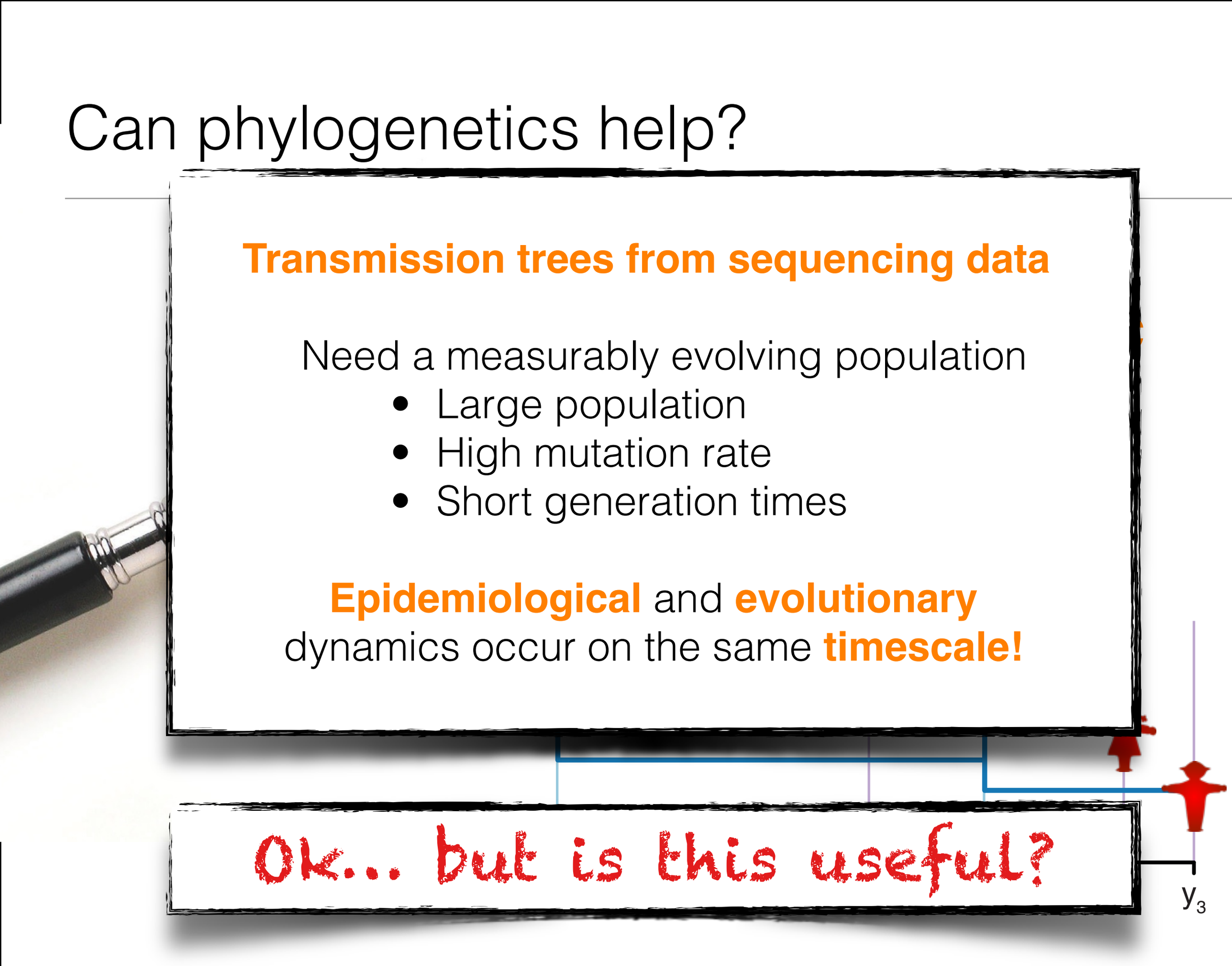
Can phylogenetics help?

Transmission trees from sequencing data

Need a measurably evolving population

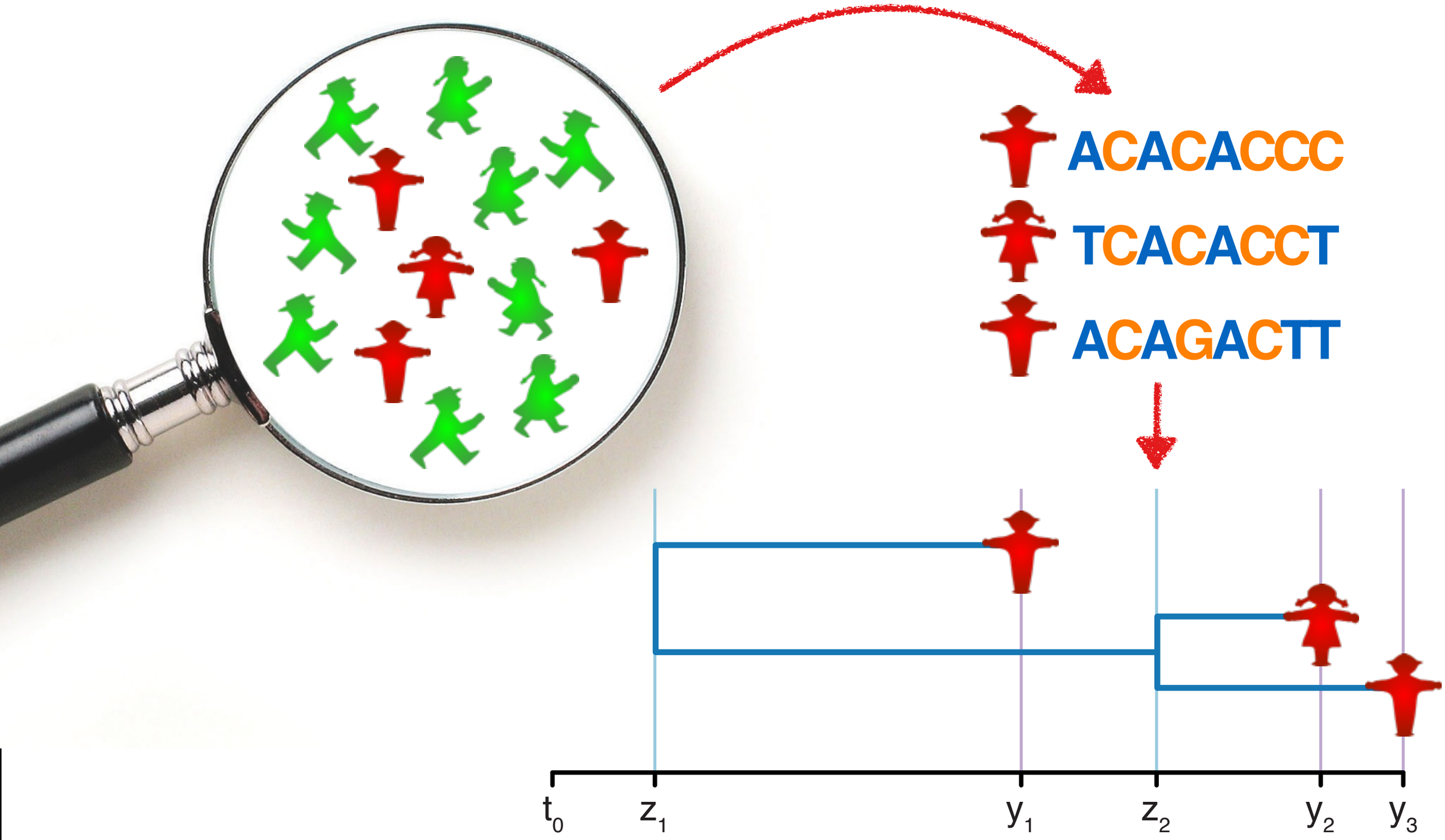
- Large population
- High mutation rate
- Short generation times

Epidemiological and **evolutionary** dynamics occur on the same **timescale!**



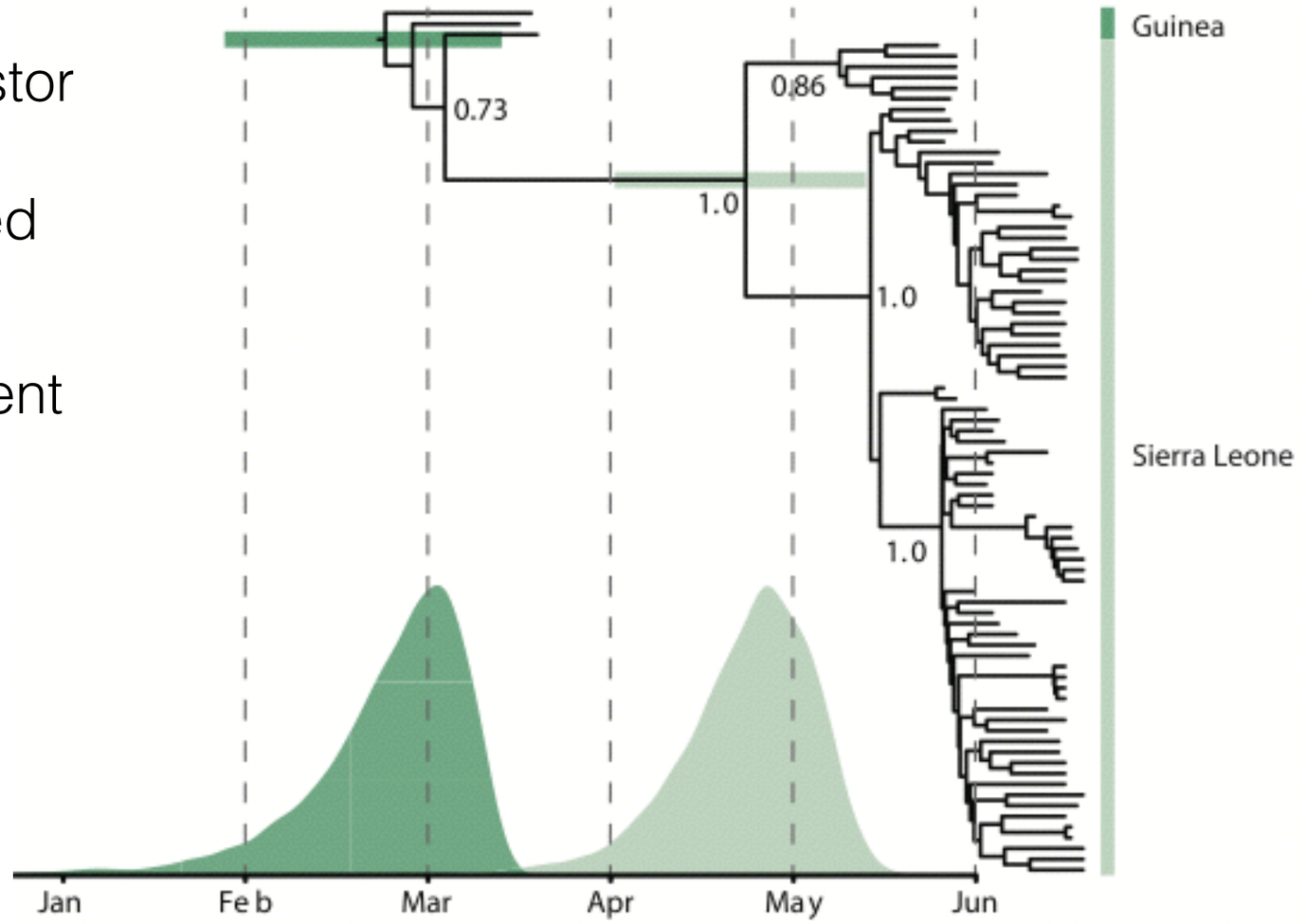
Ok... but is this useful?

Can phylogenetics help?



Tracing epidemic origins: Ebola in West Africa

- Most recent common ancestor (MRCA) in February 2014 (2 months after first recorded case)
- Likely only one zoonotic event
- Similar techniques used to date the origin of the HIV pandemic and identify the zoonotic host



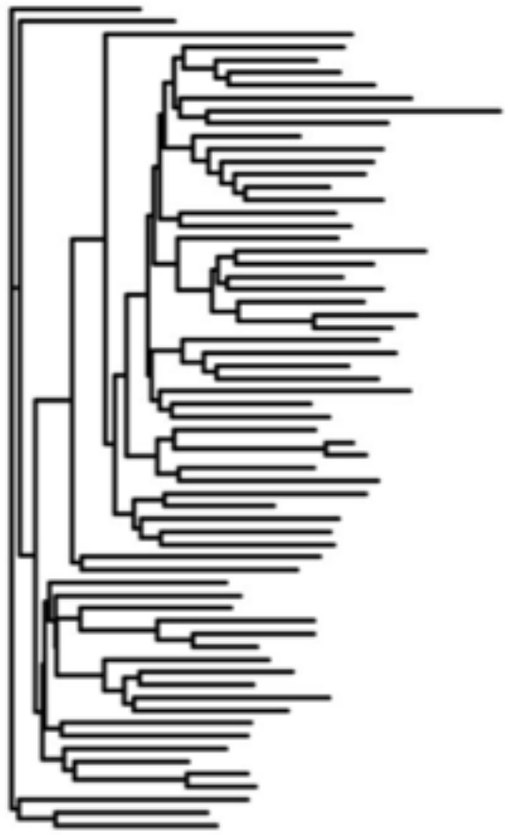
Gire *et al.* **Science 2014**

Transmission trees from sequencing data

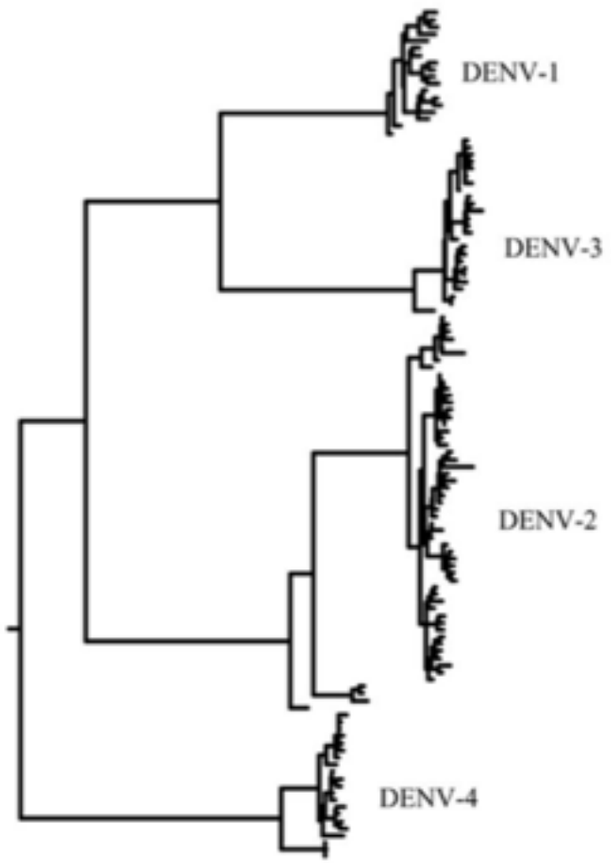
Measles

Influenza A

HCV

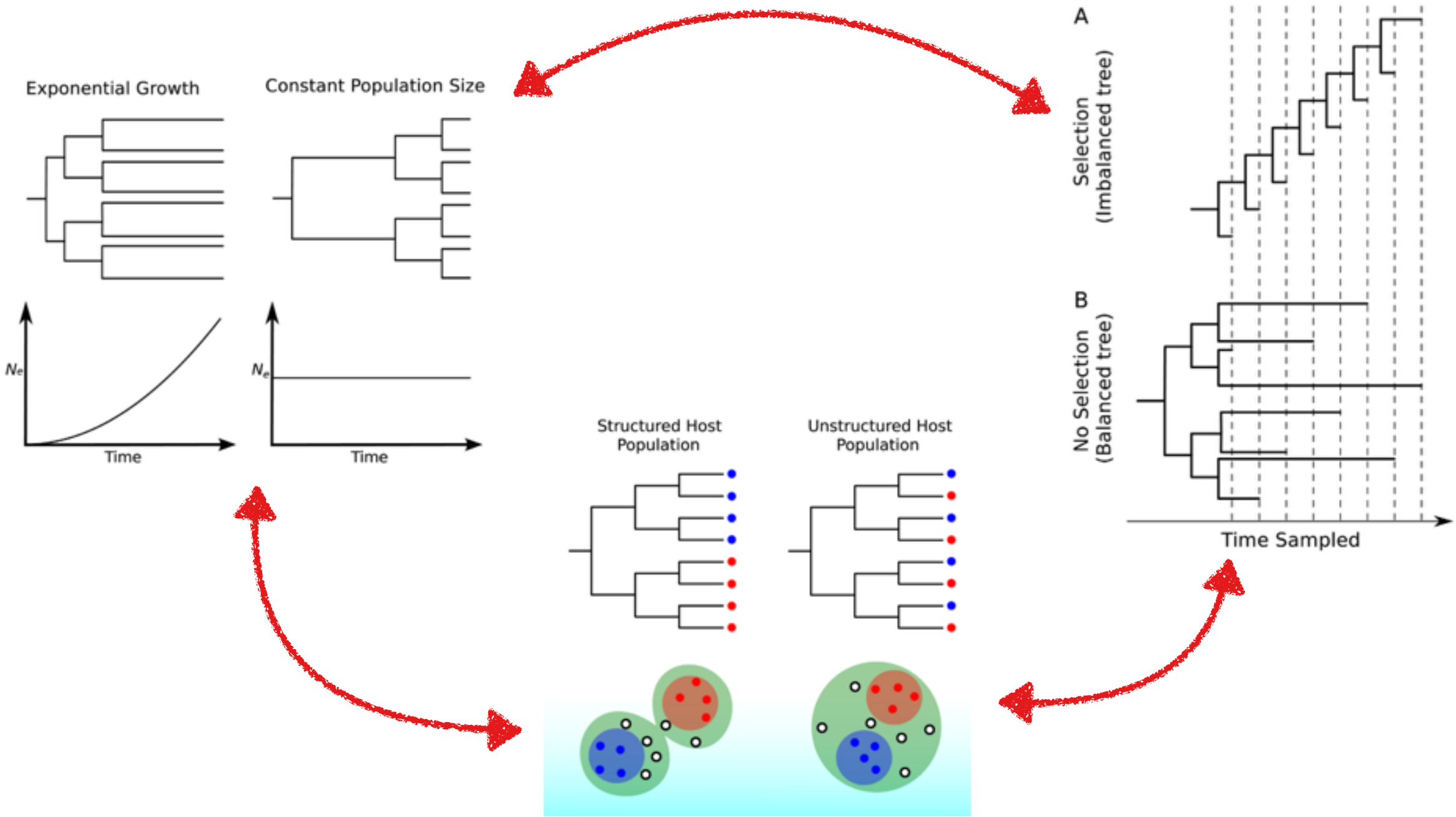


Dengue virus



Why do the trees look so different?

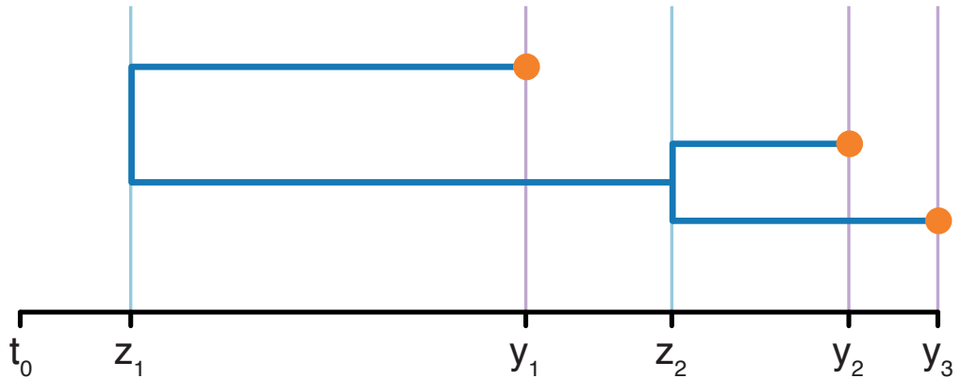
Genomes contain a signature of the epidemiological dynamics

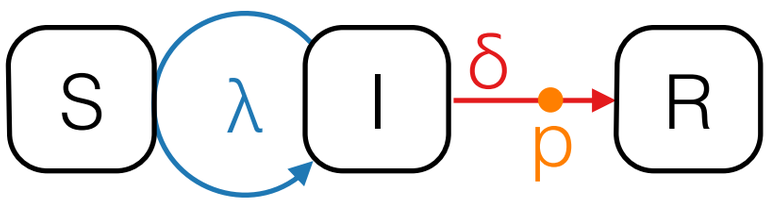


What is phylodynamics?

Phylogenetics

- State of process
- Transmission tree

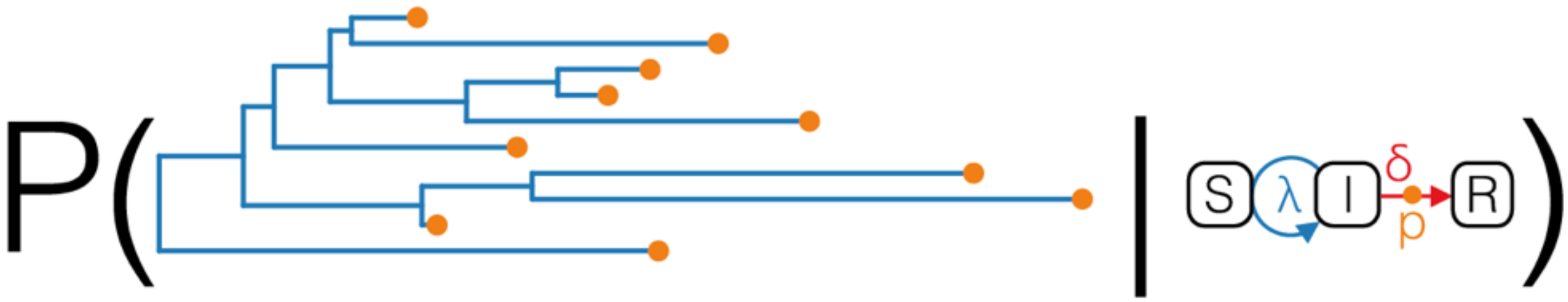


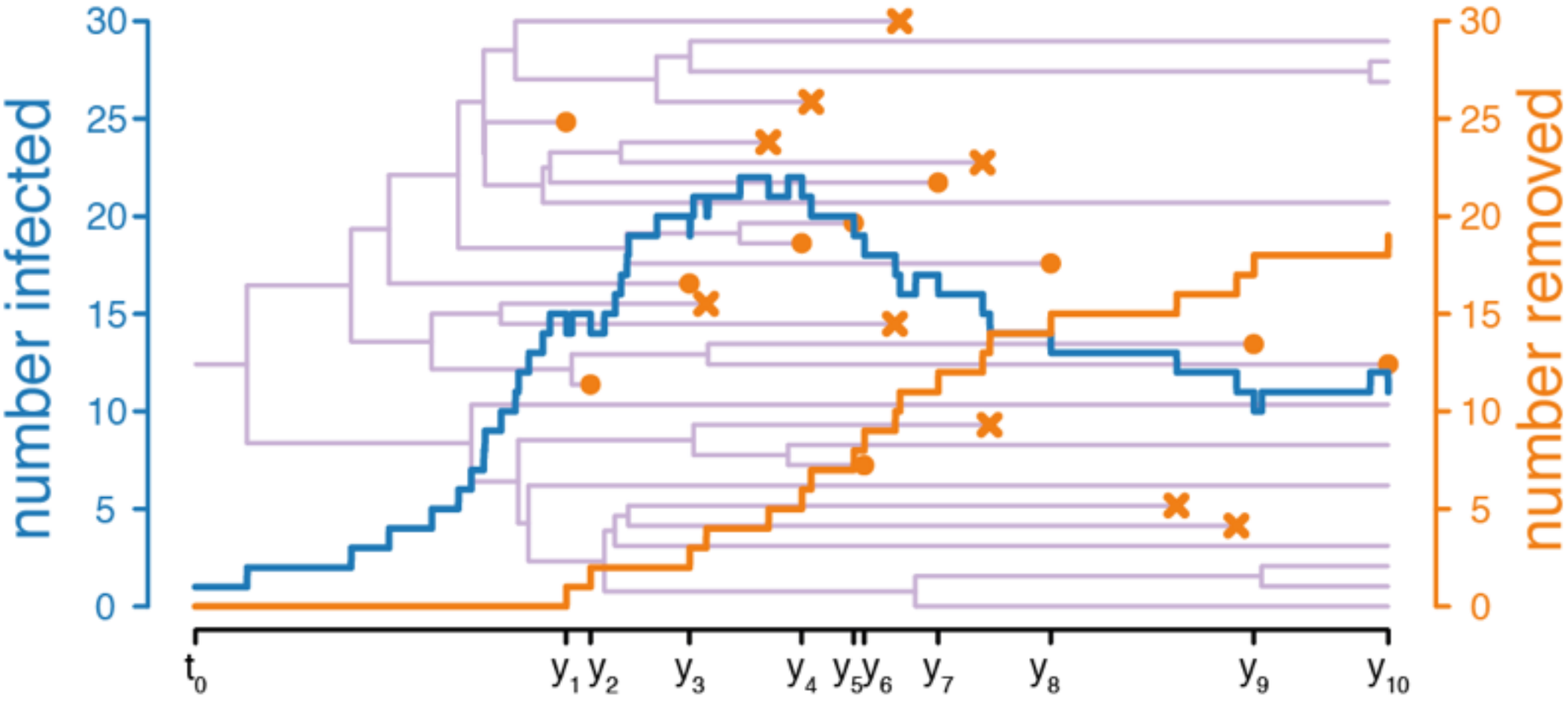


Phylodynamics

- Dynamics of process
- Epidemiological parameters

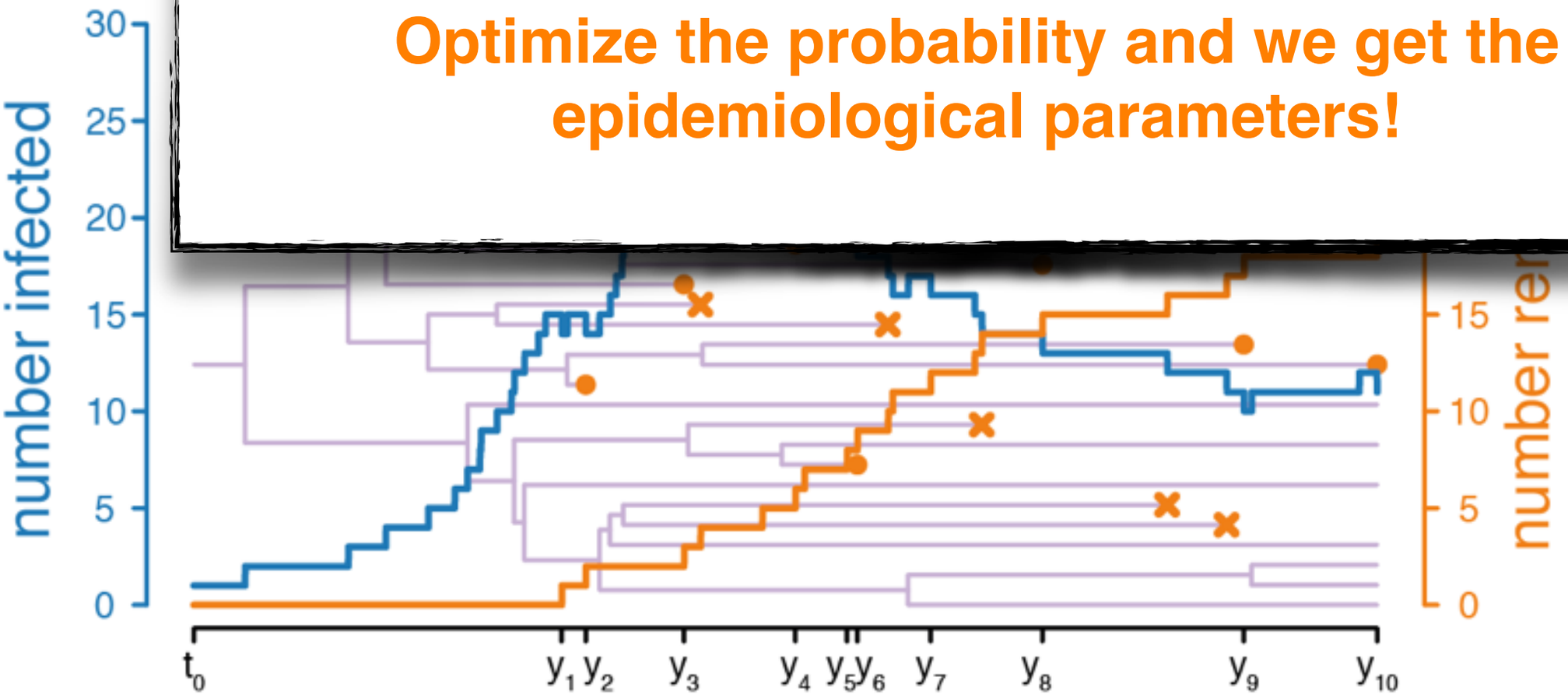
A framework for phylodynamic inference





A framework for phylodynamic inference

Optimize the probability and we get the epidemiological parameters!



A framework for phylodynamic inference

Step 1: Infer sampled transmission tree

Input data:

ACACACCC
TCACACCT
TCAGACTT
ACAGACTT

Sequence alignment

Model parameters:

Tree

Substitution model

Molecular clock model

Optimize:

$P(\text{alignment} \mid \text{tree}, \text{substitution model}, \text{clock model})$

Find model parameters that best describe the evolution of sequences along a tree

Step 2: Infer dynamics that led to the tree

Input data:

Tree

(Realisation of a stochastic process)

Model parameters:

Demographic model

(Model describes dynamics of pathogen transmission)

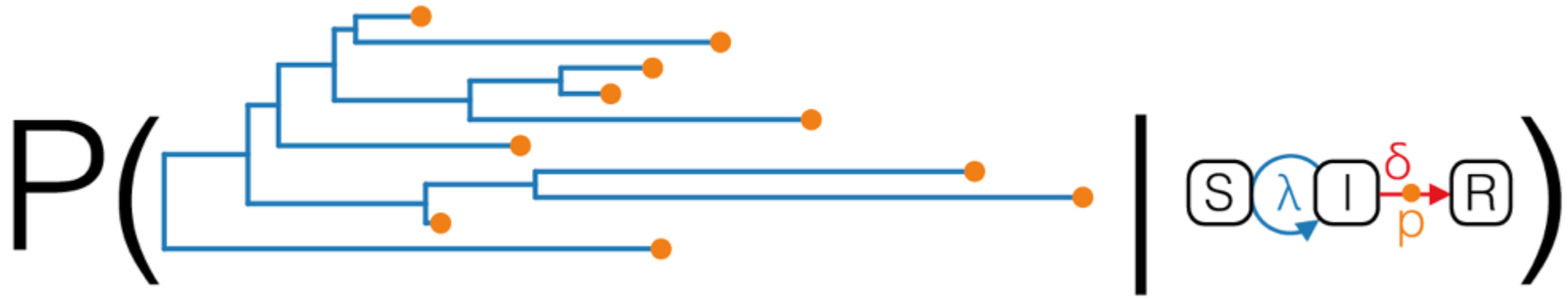
Optimize:

$P(\text{tree} \mid \text{demographic model})$

Find model parameters that best describe how the tree grows during an epidemic

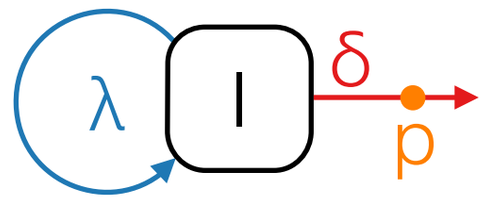
Major assumption: Patients do not transmit after they are sampled!

Birth-death models (forward in time)

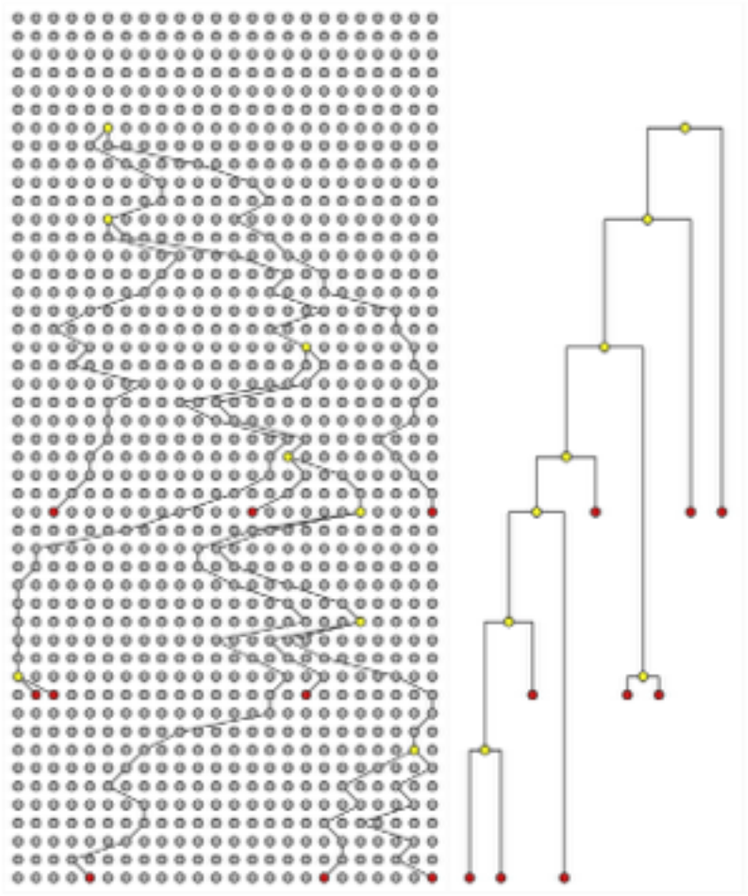


- Directly calculate the probability of observing a sampled tree given a model
- Computationally expensive for more complex models
- Cannot calculate probability in general

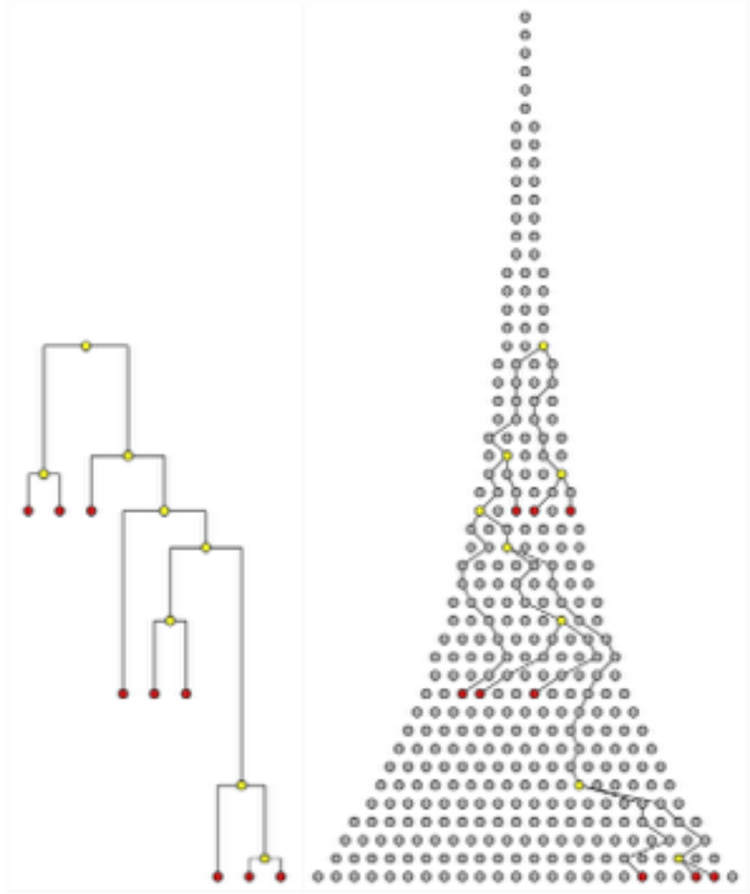
- Usually stick with simple models:



Coalescent approximation (backward in time)



- Relates the shape of the tree to population size
- Assume a random sample drawn from a population
- Calculate the probability of two lineages coalescing in a given time under a Wright-Fisher process
- Calculate the probability of observing the coalescent times in the sampled tree
- Infer effective population size (N_e) or growth rate



- Originally assumed a constant population size
- Generalized to any deterministically changing population size
  (eg. exponential growth, SIR dynamics)

Summary: Phylodynamic inference framework

Epidemiological and evolutionary dynamics on the same timescale

↓

Pathogen genomes contain signature of epidemic spread

↓

Leaves pattern on resulting phylogeny (sampled transmission chain)

↓

Extract epidemiological parameters from phylogeny

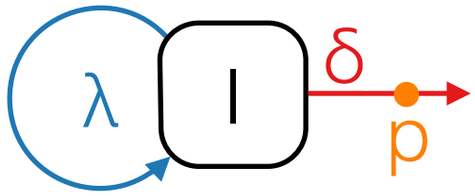
↙ ↘

Birth-death models Coalescent models

Summary: Phylodynamic inference framework

Epidemiological parameters example

Constant rate birth-death process



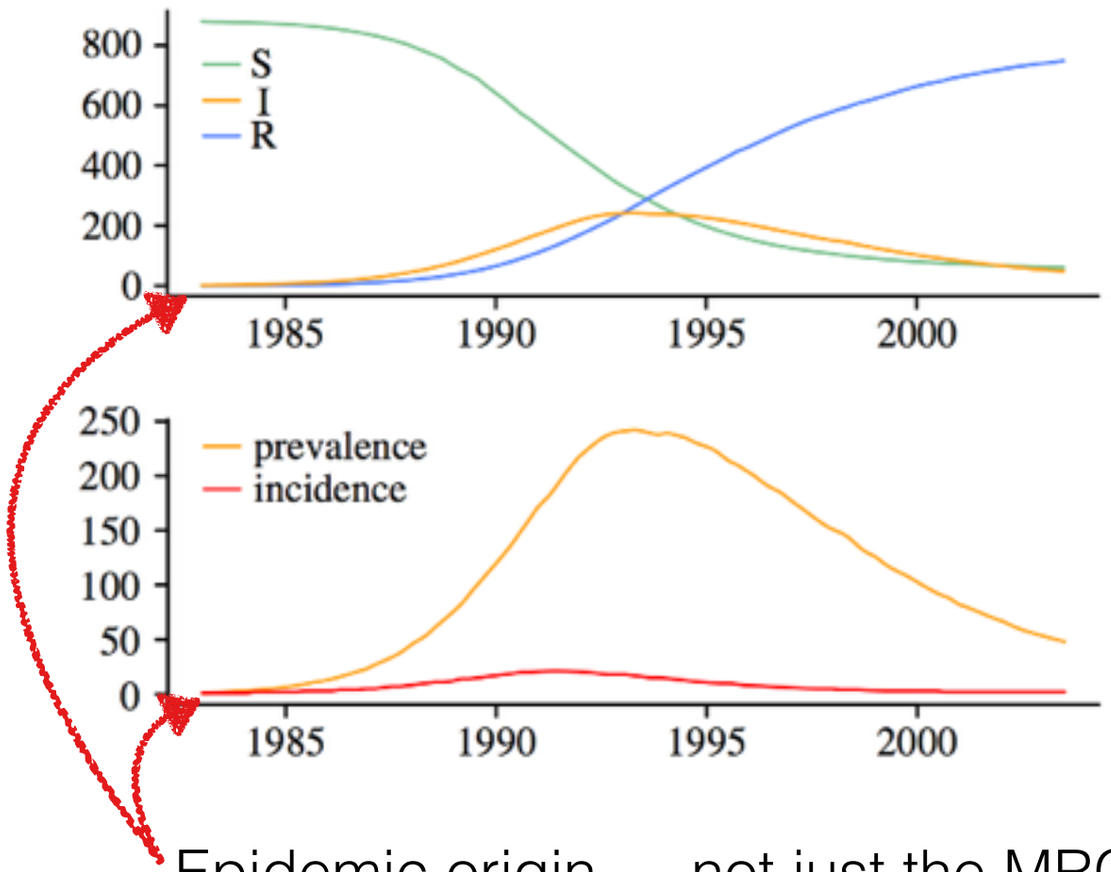
- λ — infection rate
- δ — becoming-noninfectious rate
- p — sampling probability

- $R_0 = R_e = \lambda/\delta$
- Infectious period = $1/\delta$
- Epidemic origin: Time when I = 0
- Number infected: Growing exponentially at rate λ-δ
- …

Birth-death models Coalescent models

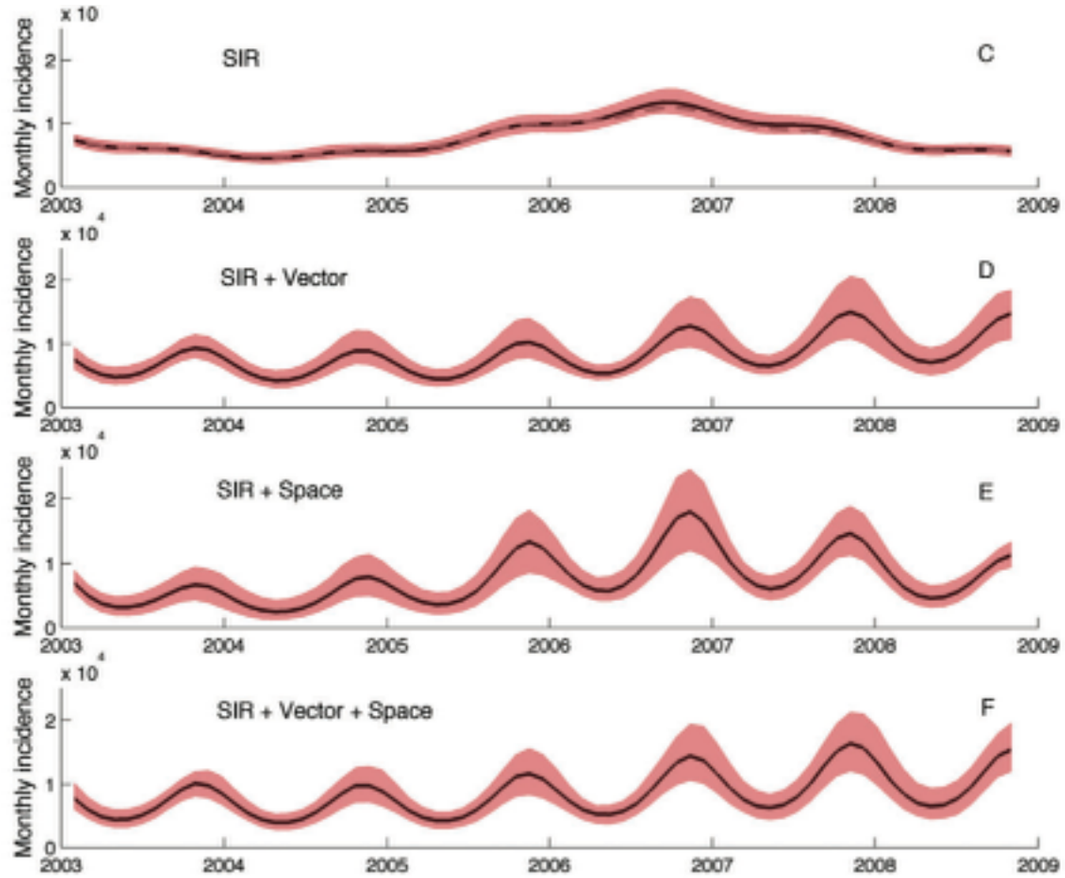
Mechanistic epidemiological models

Birth-death — HIV-1



Epidemic origin — not just the MRCA!

Coalescent — DENV-1



Easier to handle more complex mechanistic dynamics in a coalescent framework!

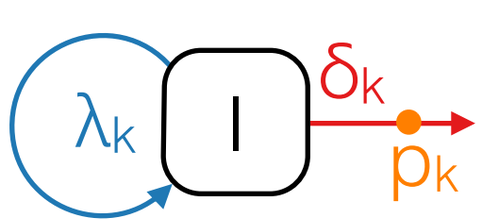
Kühnert *et al.* **J R Soc Interface 2014**

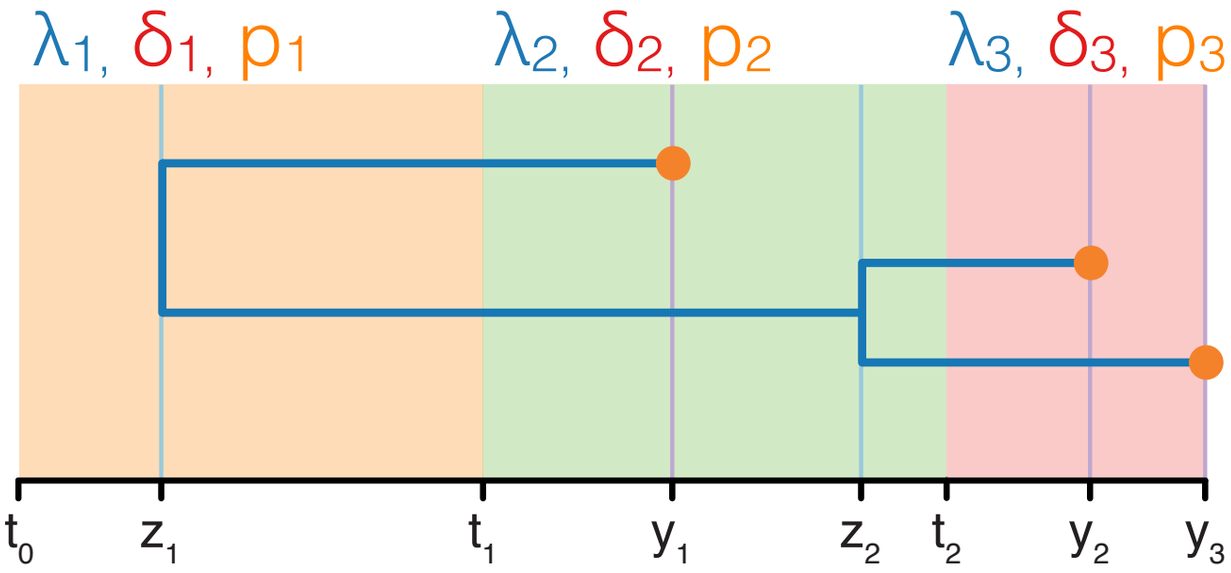
Rasmussen *et al.* **MBE 2014**

Nonparametric models: Skyline approximation

- Mechanistic models attempt to exactly describe the dynamics of the epidemic
- Not always possible to derive an expression for the likelihood under complex dynamics
- Exact dynamics are not always known

Solution: Piecewise constant approximation



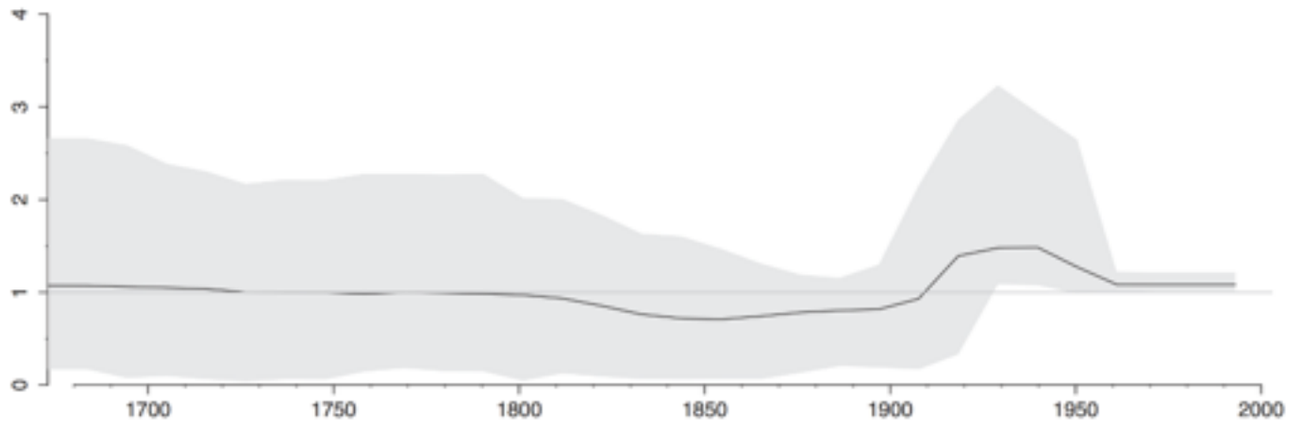


- **Birth-death skyline:**
  Rates change over time
- **Coalescent skyline:**
  N_e changes over time
- Can approximate any type of dynamics (if population structure not important)
- Need more data
  (no prior information on dynamics)

Skyline approximation: HCV in Egypt

Coalescent Skyline — Changes in effective population size, N_e

Birth-death Skyline — Changes in effective reproduction number, R_e

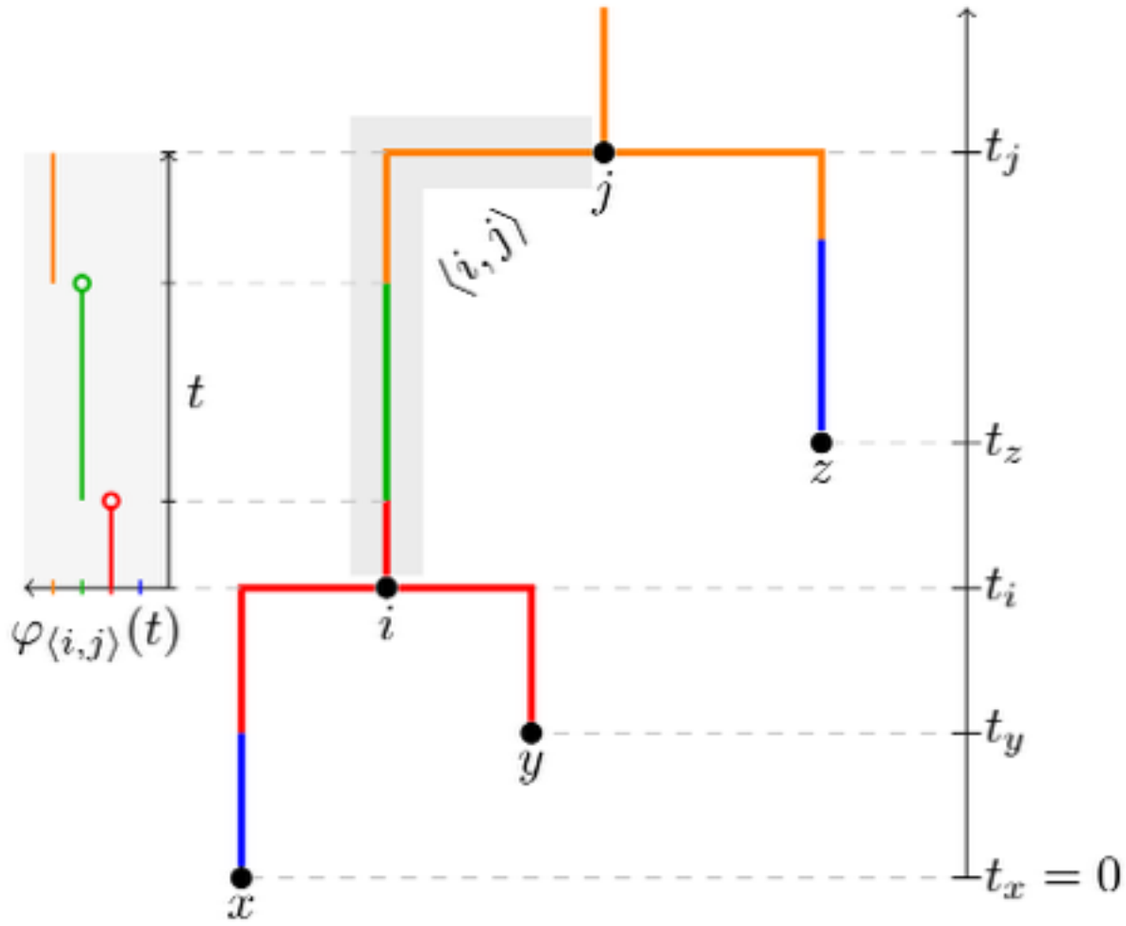


Large increase in N_e/R_e around 1950s due to antischistosomal injections with contaminated needles!

Stadler *et al.* **PNAS 2012**

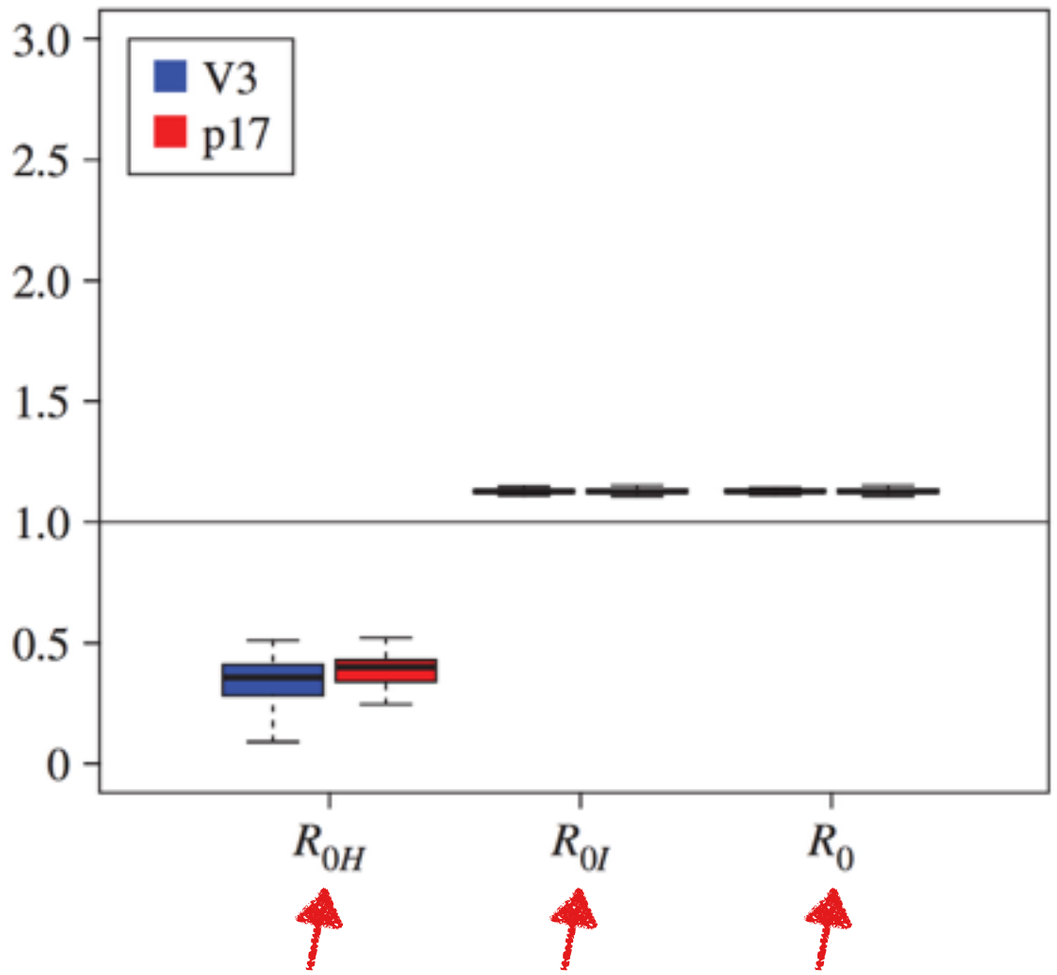
Population structure

HIV-1 in Latvia



- Explicitly model different types of subpopulations
- Different types may be associated with different epidemiological parameters
- Can also model migration between types

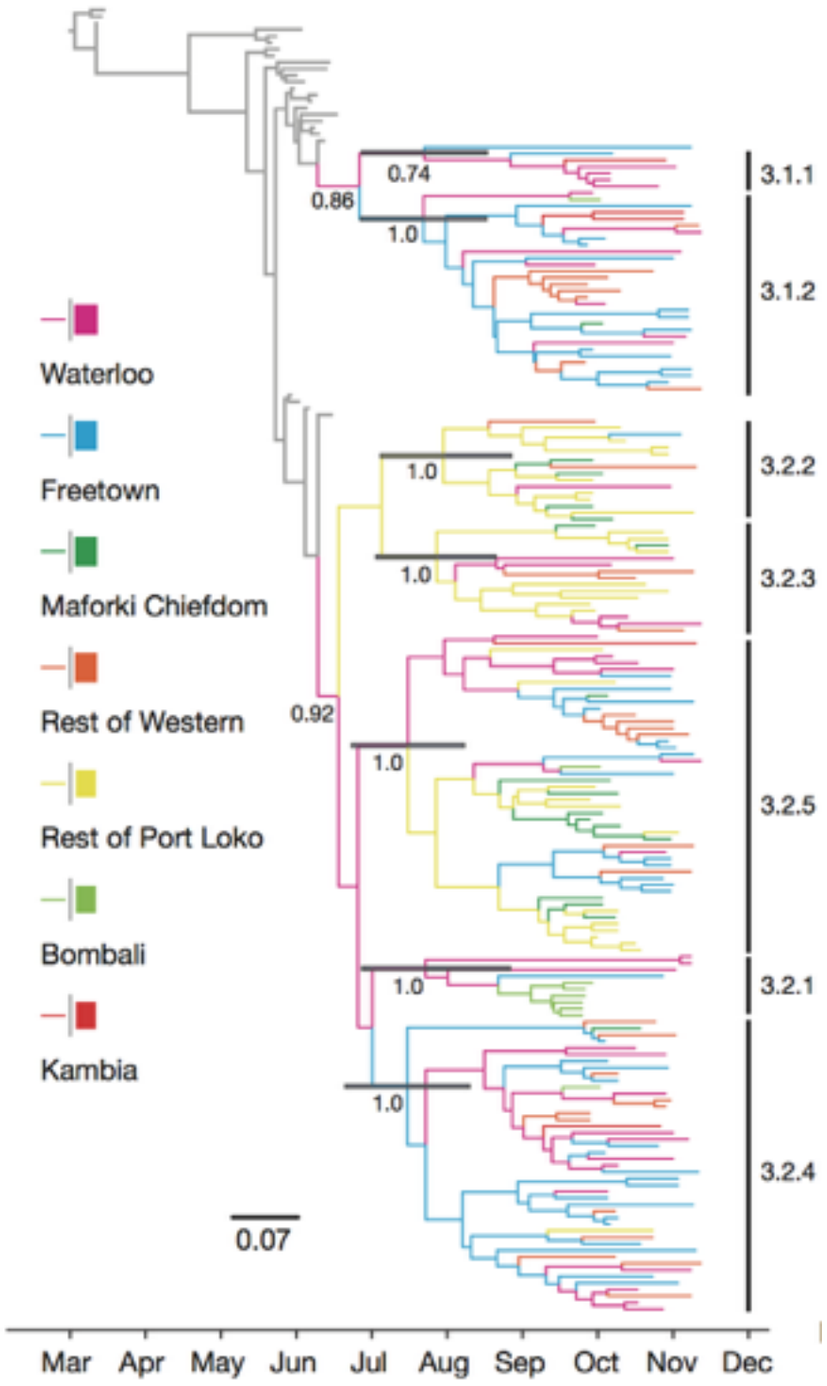
Vaughan *et al.* **Bioinformatics 2015**

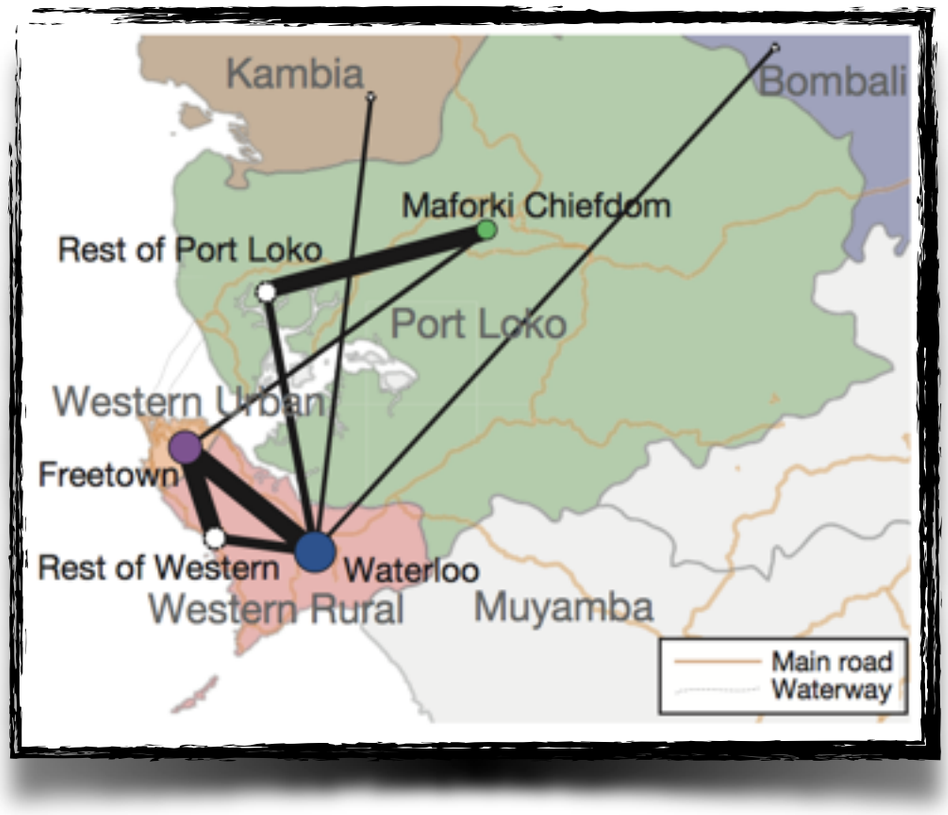


- Injecting drug users have $R_0 > 1$
- This is keeping the epidemic alive!

Stadler and Bonhoeffer *et al.* **Proc R Soc B 2013**

Phylogeography: Ebola in West Africa



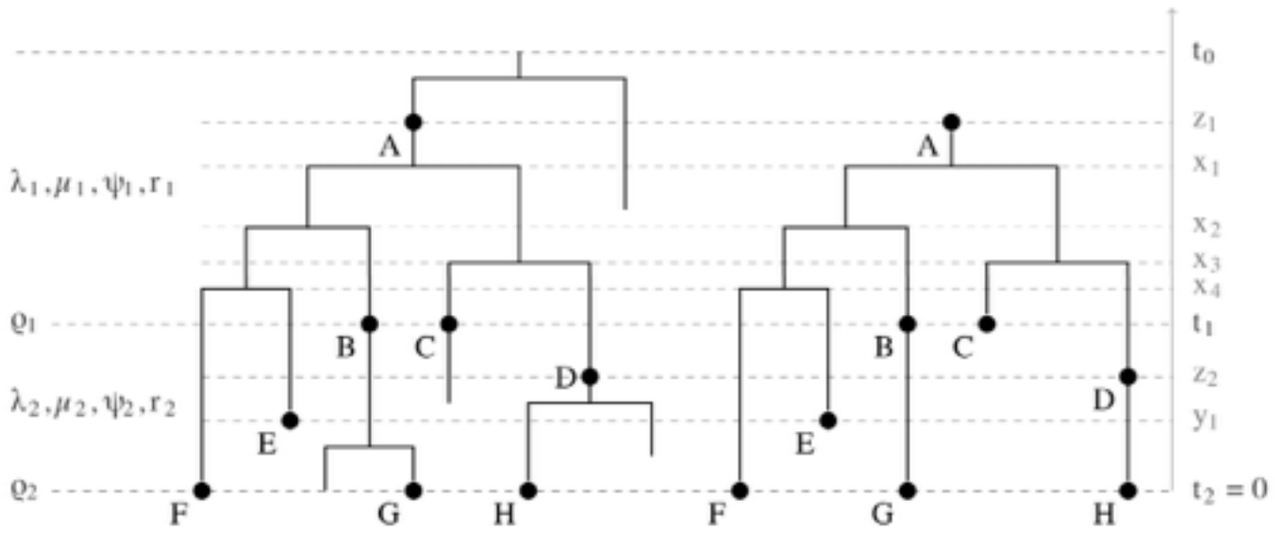


- Infer how the virus spread across the region
- Infer the most probable ancestral states
- Infer migration rates between different regions

Tong *et al.* **Science 2015**

Sampled ancestors

- Sometimes individuals keep on transmitting after they are sampled (eg. HIV-1, Ebola virus)
- Can relax this assumption for birth-death models



Gavryushkina *et al.* **PLoS Comput Biol 2014**

Samp

- Someti
  after th
  (eg. HI
- Can re
  models

$\lambda_1, \mu_1, \psi_1, r_1$

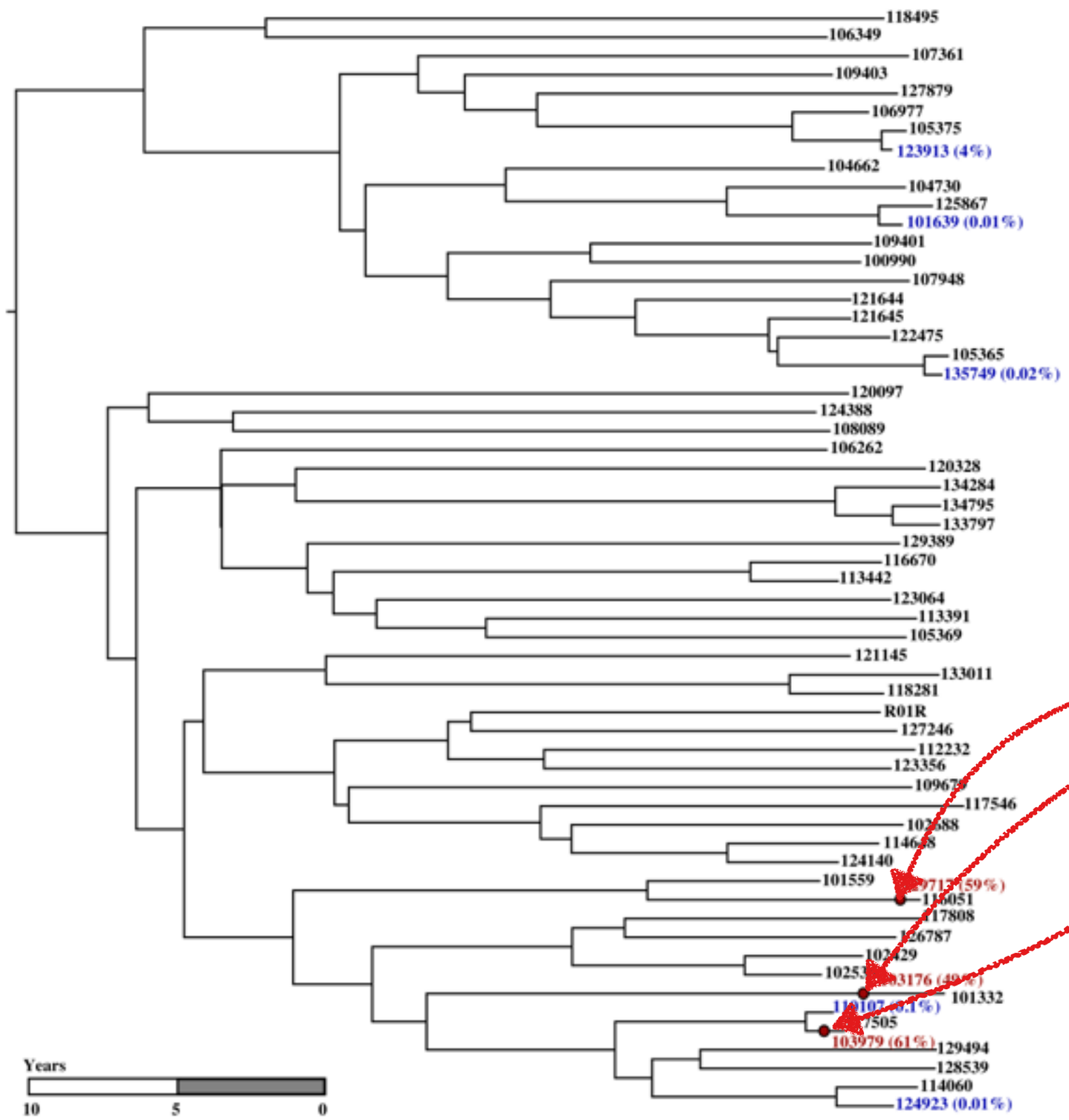
ϱ_1

$\lambda_2, \mu_2, \psi_2, r_2$

ϱ_2

F

HIV-1



Sampled ancestors!

Gavryushkina *et al.* **PLoS Comput Biol 2014**

Birth-death

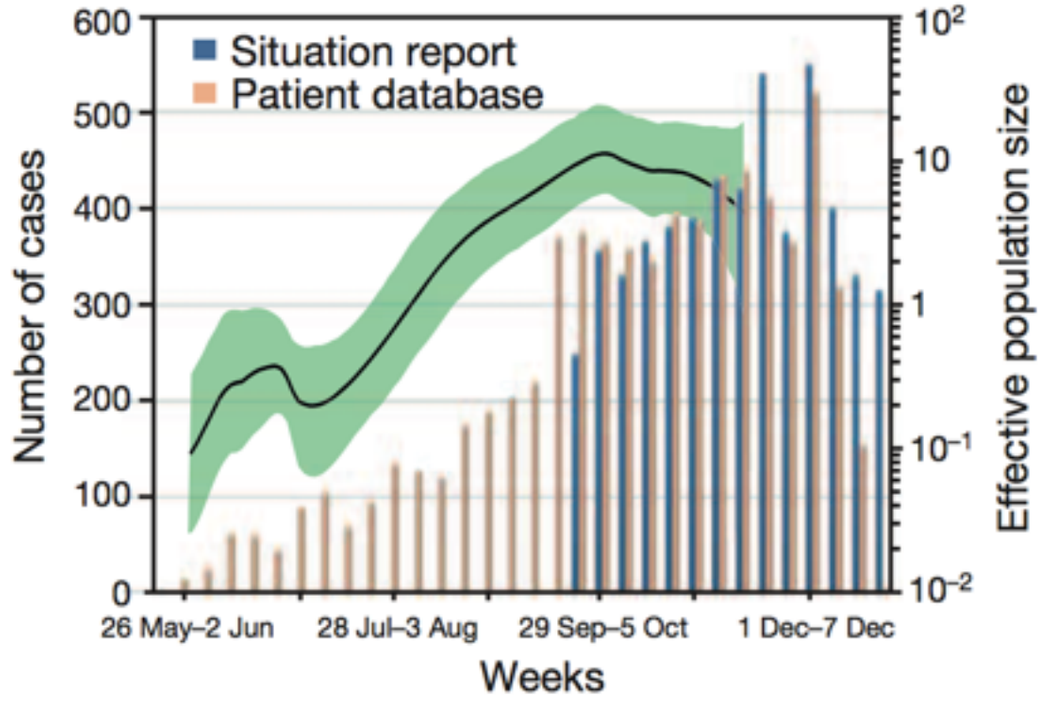
- Better treatment of stochasticity 😄
- Good for large sampling proportion and small population sizes 😄
- More computationally expensive 😖
- Cannot easily calculate the likelihood for complex models 😰

Coalescent

- Not as good at modelling stochasticity 😔
- Assumes a small sampling proportion 😖
- Computationally relatively fast 😄
- Easier to derive coalescent times under complex models 😄

But both models should give similar results most of the time…

Section 2: Parameter inference



Louis du Plessis

Reminder: 2-step method for phylodynamic inference


genetic sequences


genealogy


demographic model


substitution model


molecular clock model

Step 1: Infer sampled transmission tree

Find model parameters that best describe the evolution of sequences along a tree

Step 2: Infer dynamics that led to the tree

Find model parameters that best describe how the tree grows during an epidemic

Tree and its dynamics are not always independent so a 1-step method would be better…

Maximum likelihood

- Find that maximises:

$$P(\text{tree} \mid \text{model})$$

- Point estimates
- May get stuck in local optima

MCMC

- Find distribution of given 
- Posterior of parameters:

$$P(\text{model} \mid \text{tree})$$

- Explore entire likelihood surface

Prior information

- Can include prior information in MCMC:

$$P(\text{model}\ \text{parameters}\ \text{prior}) = P(\text{model})P(\text{parameters})P(\text{prior})$$

- Priors can bias results if too strong

Posterior for MCMC inference

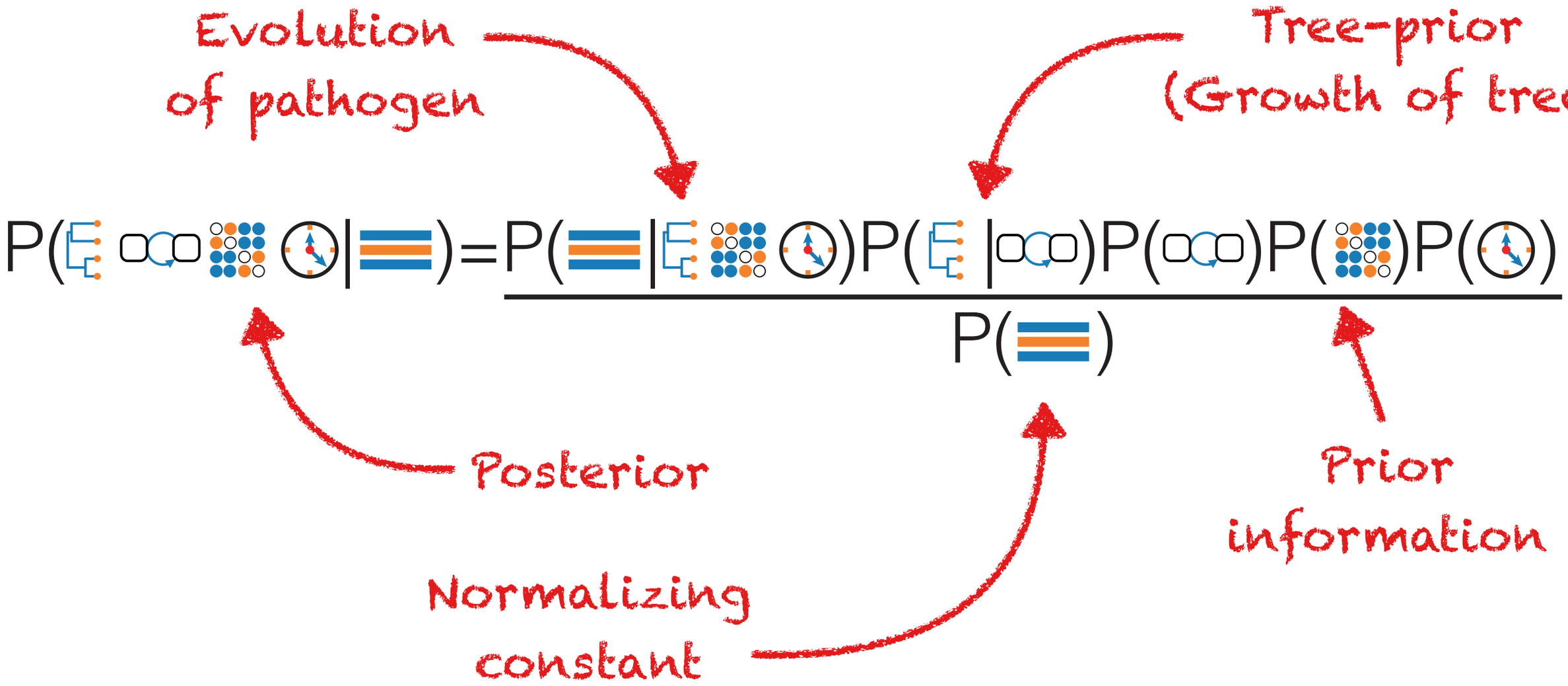
genetic sequences

genealogy

demographic model

substitution model

molecular clock model



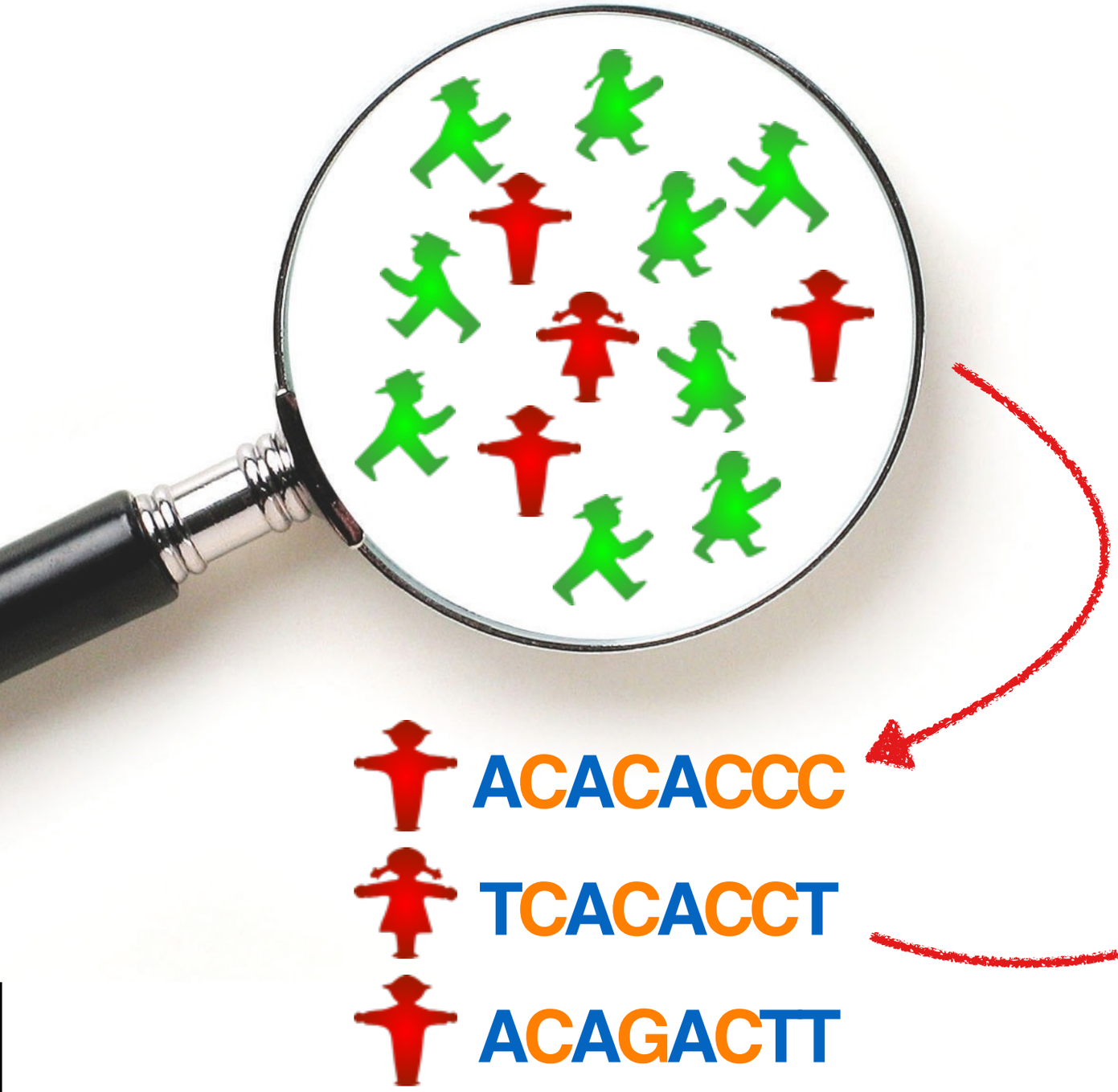
One tree to rule them all?

- As with other parameters the data may support multiple trees equally well
- Thus we should look at all trees in order to take phylogenetic uncertainty into account
- If we are only interested in the epidemiological parameters we can integrate out the tree (but this is very difficult in an ML framework)
- Naturally integrate out the tree in an MCMC framework by inferring the distribution of trees

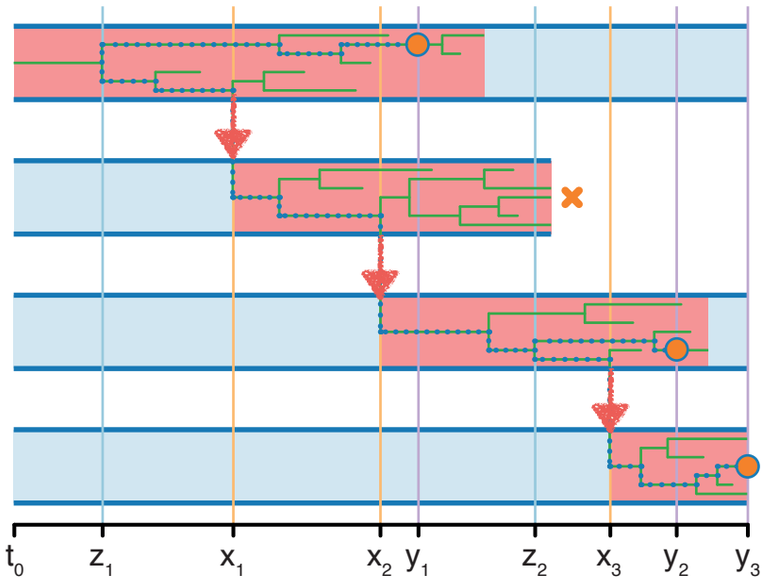
Epidemiological parameters directly from sequencing data!

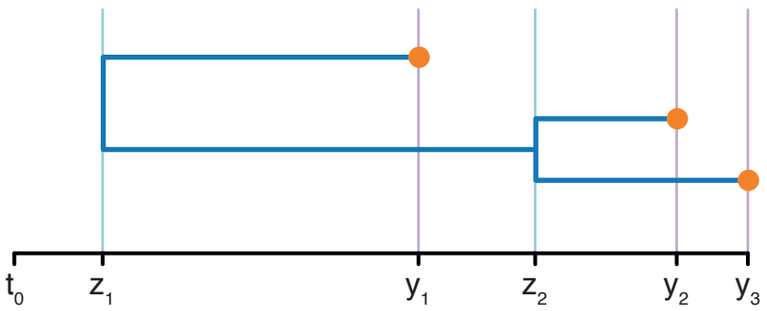
genetic sequences | genealogy | demographic model | substitution model | molecular clock model

Sampled transmission trees revisited



- No information on who infected whom
- Transmission events usually more recent than the MRCAs of the lineages

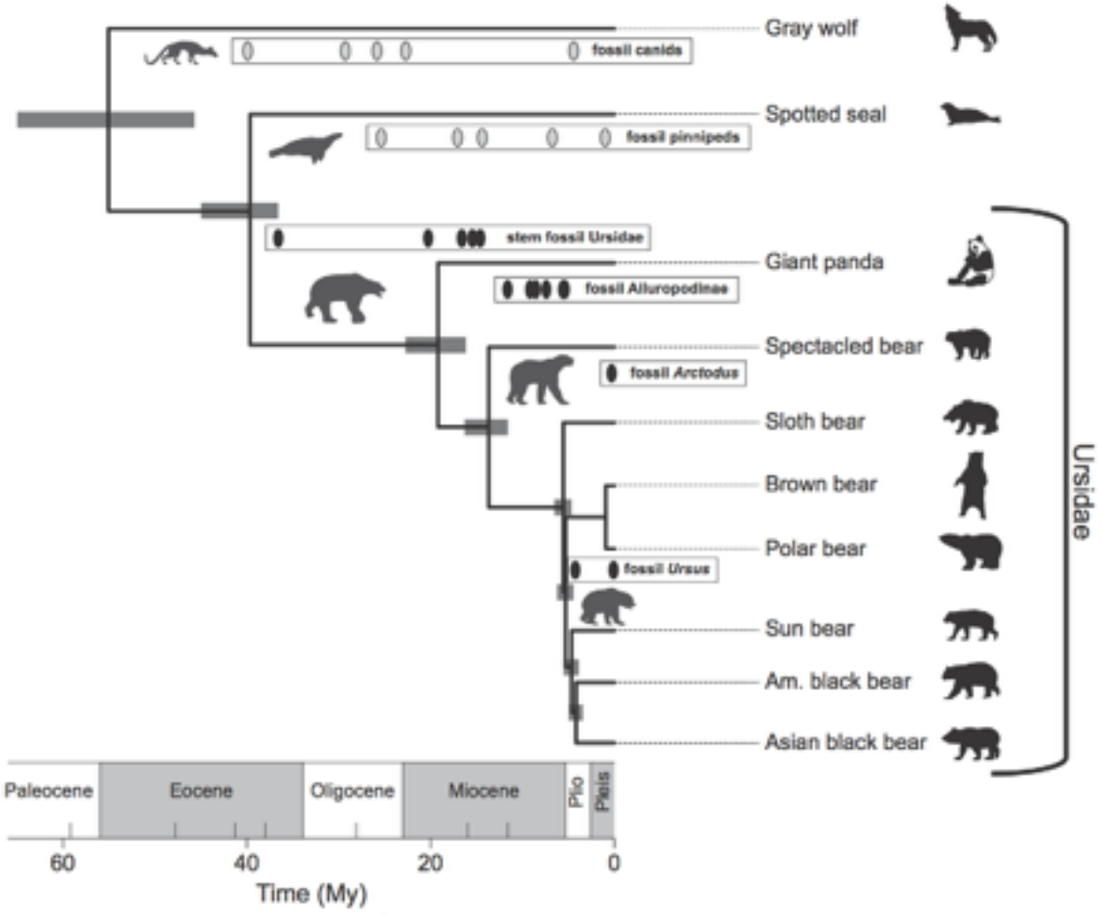




Phylodynamics is not just about viruses!

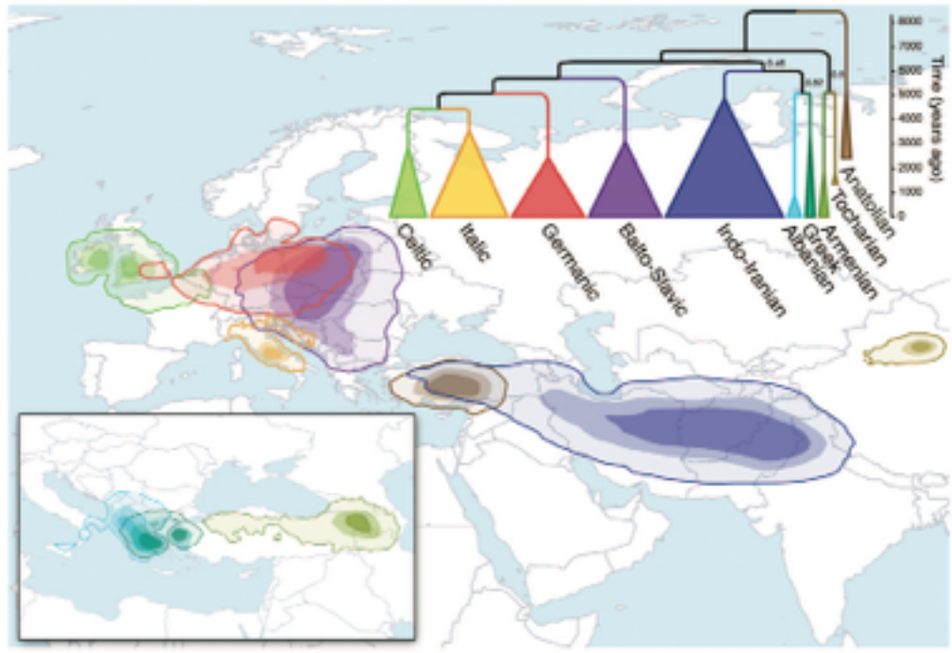
Phylodynamic methods are applicable to any infectious disease with enough phylogenetic diversity…

…or macro-evolution…



Heath *et al.* **PNAS 2014**

…or language evolution…

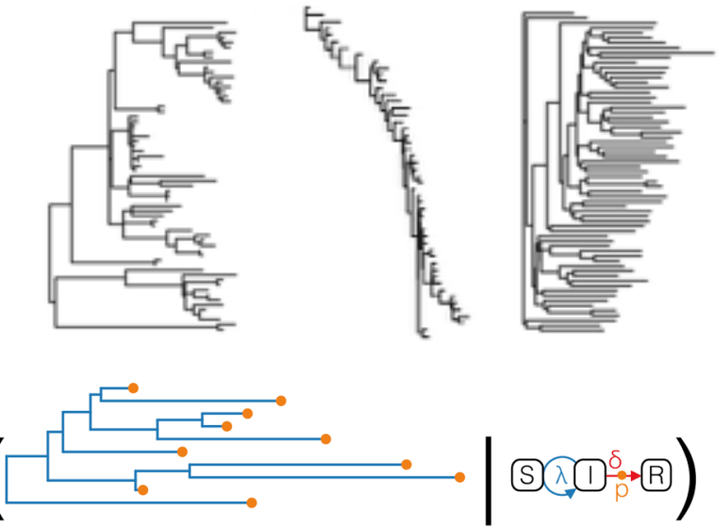


Bouckaert *et al.* **Science 2012**

…and more!

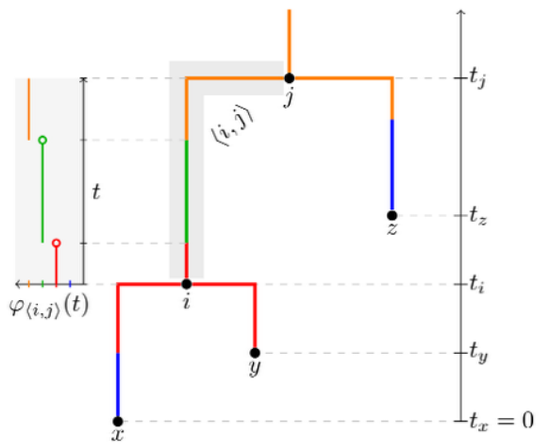
Summary

- Epidemiological dynamics leaves a signal on the branching pattern in the phylogeny



- Phylodynamic models can extract this signal by calculating the probability of the observed phylogeny conditioned on a demographic model

- Models can be specified in a birth-death or coalescent framework and may be more or less complex based on the dynamics that are to be modelled



- Using MCMC we can infer the posterior distributions of epidemiological parameters while accounting for phylogenetic uncertainty

$$P(\cdot|\cdot)=\frac{P(\cdot|\cdot)P(\cdot|\cdot)P(\cdot)P(\cdot)P(\cdot)}{P(\cdot)}$$

Grenfell *et al.* **Science 2004**
Vaughan *et al.* **Bioinformatics 2015**