# Phylogenetic insights into infectious disease epidemiology

Molecular Epidemiology of Infectious Diseases

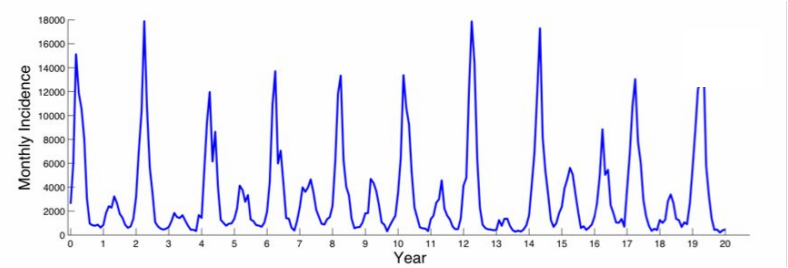
Lecture 1

January 6th, 2020

Genomic data has given us new power to track the spread of infectious pathogens

# Revealing the source of infections

Classic sources of epidemiological data are typically not informative about the sources of new infections




Pathogen genomic data can help **reveal the movement of pathogens and the source of new infections**

# The basic idea

Genetic relatedness of pathogens sampled from different hosts or environments provides us with information about possible transmission routes

# The importance of phylogenies

While there are many methods for analyzing pathogen genomic data, this lecture and most of this class will examine phylogenetic methods in molecular epidemiology.

# Why phylogenies?

1. Phylogenies can easily be related back to the epidemic dynamics of a pathogen

2. Thinking phylogenetically can help us understand how epidemic dynamics shape genetic variation in a pathogen population.

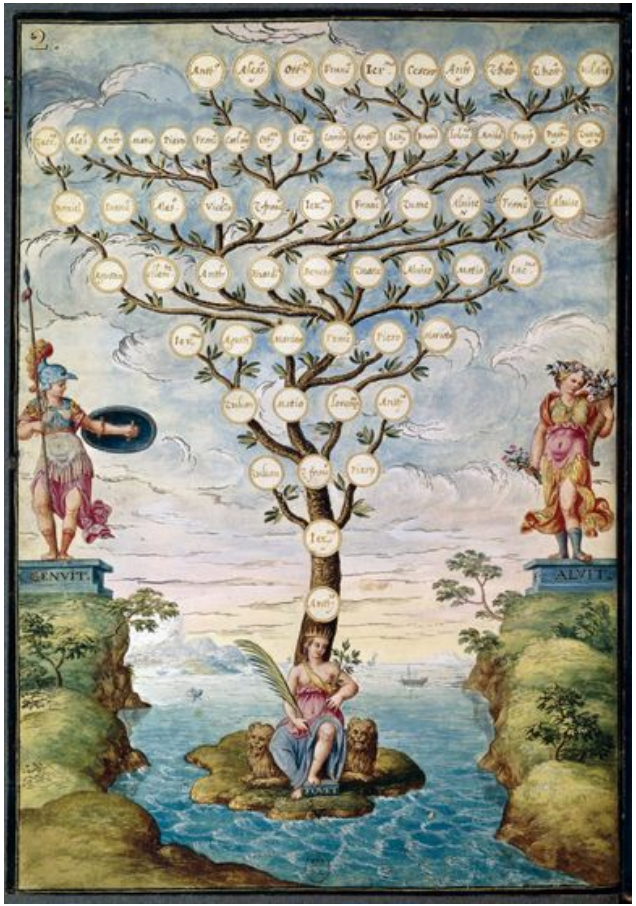
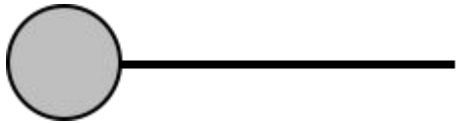


Image from *The Book of Trees* (Manuel Lima, 2014)

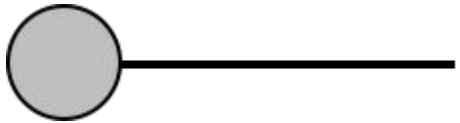
Let’s start by considering a small epidemic spreading through a host population

# A simple epidemic example

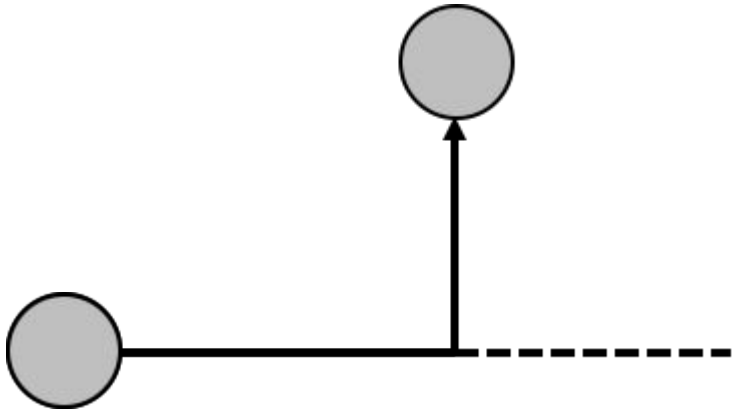


# A simple epidemic example

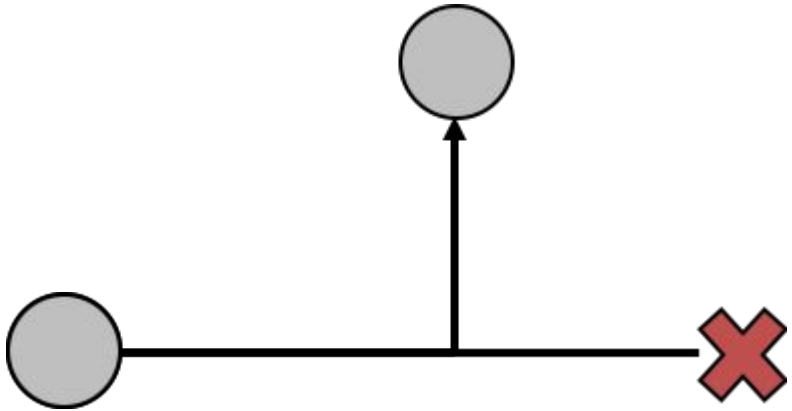
# A simple epidemic example



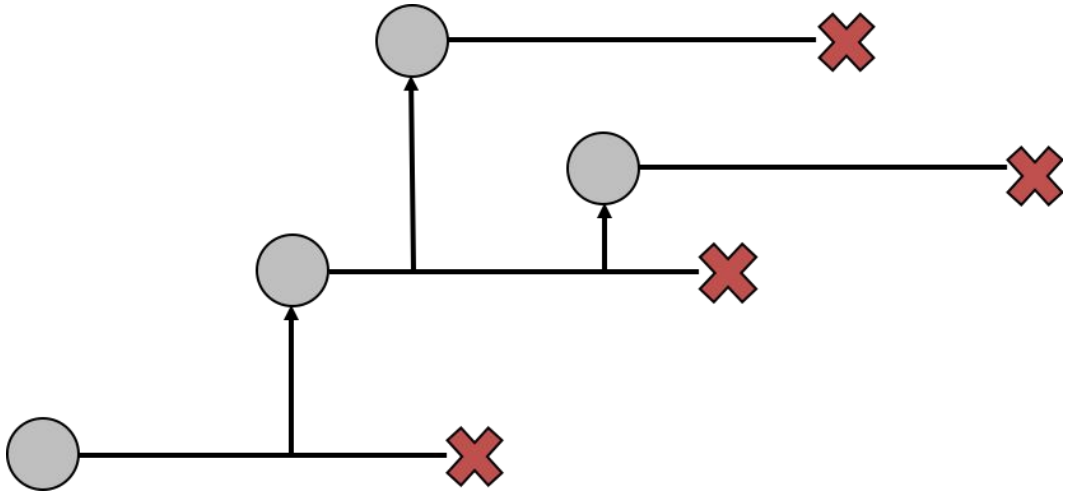
# A simple epidemic example



# A simple epidemic example

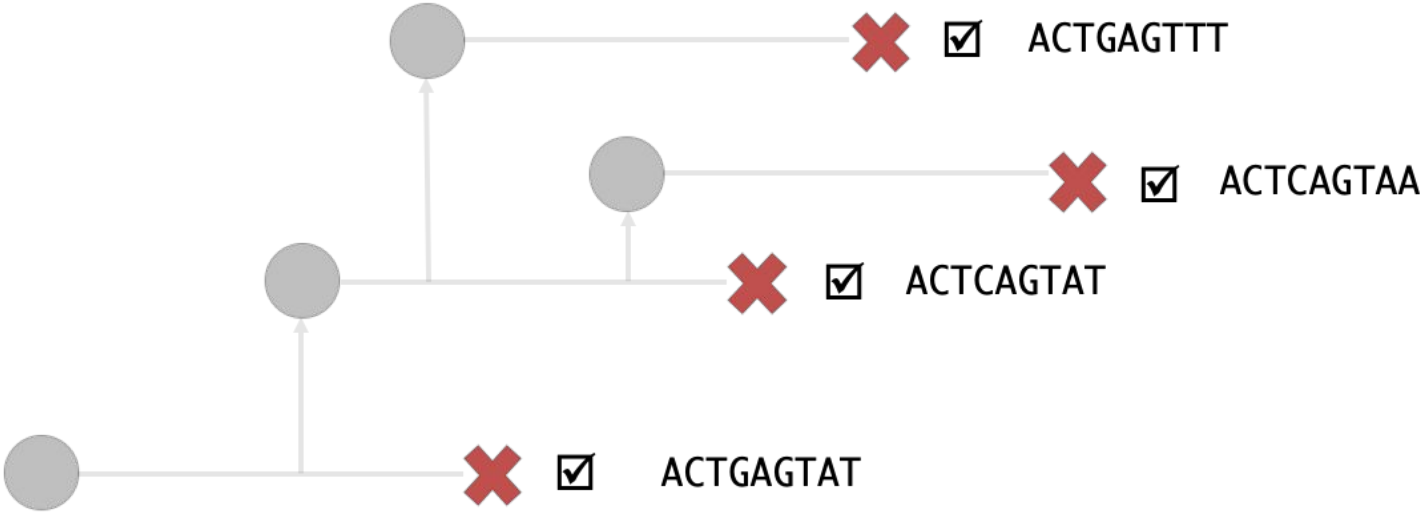


# A simple epidemic example

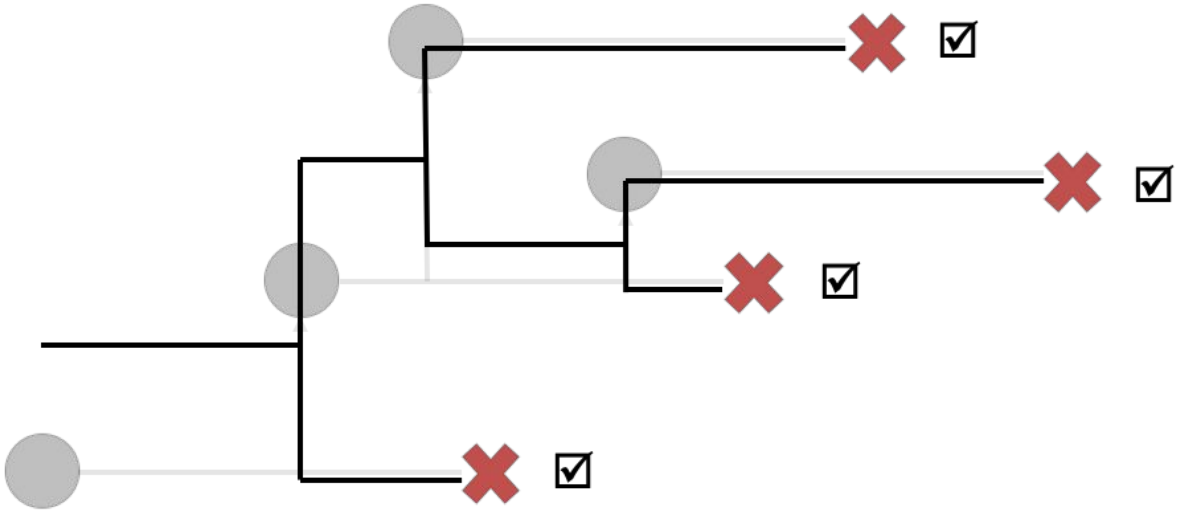


**Transmission tree**

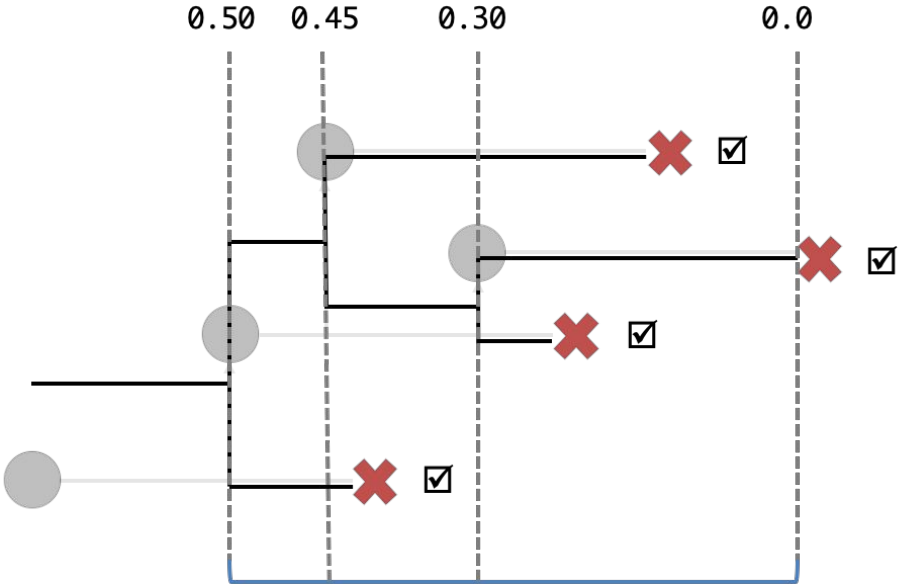
# A simple epidemic example

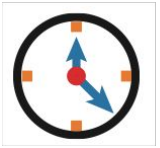


# A simple epidemic example



# A simple epidemic example



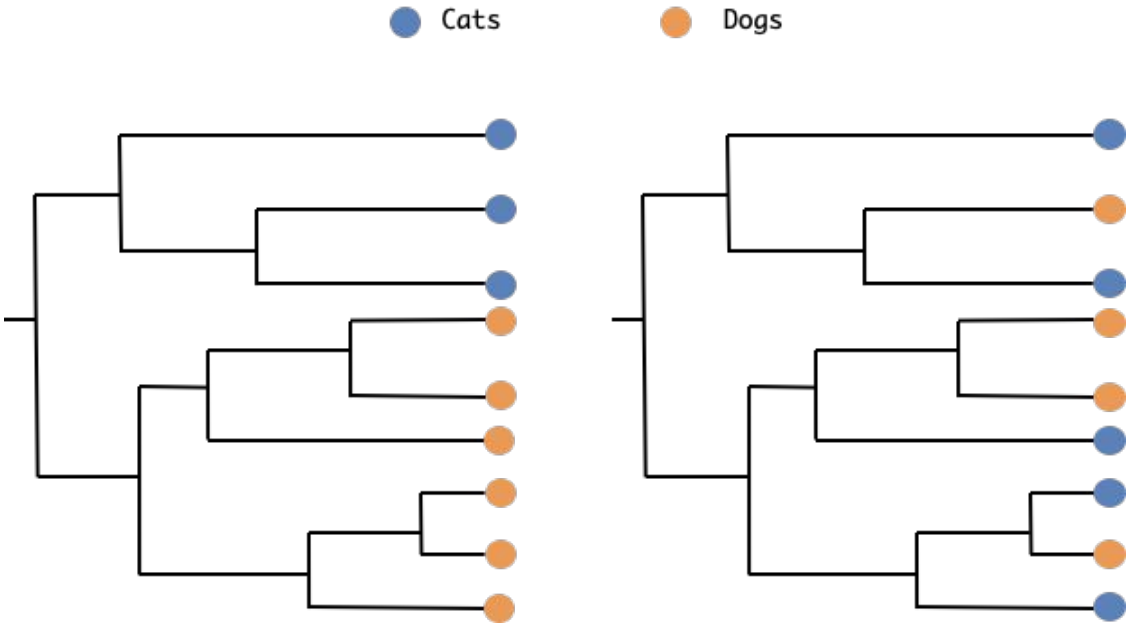



$$\text{Real time} = \frac{\text{genetic distance}}{\text{clock rate}}$$

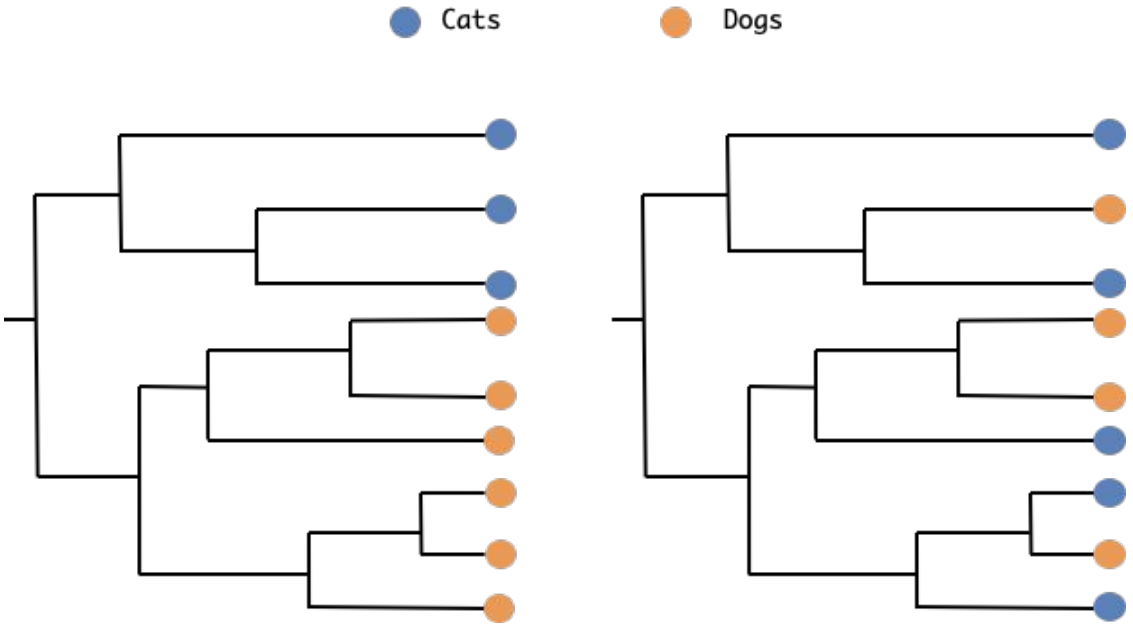
# Phylogenies can tell us about:

- Linkage and the sources of transmission
- The origins of epidemics and new strains
- Past epidemic dynamics
- Pathogen fitness and adaptation

# Phylogenetic linkage



# Phylogenetic linkage



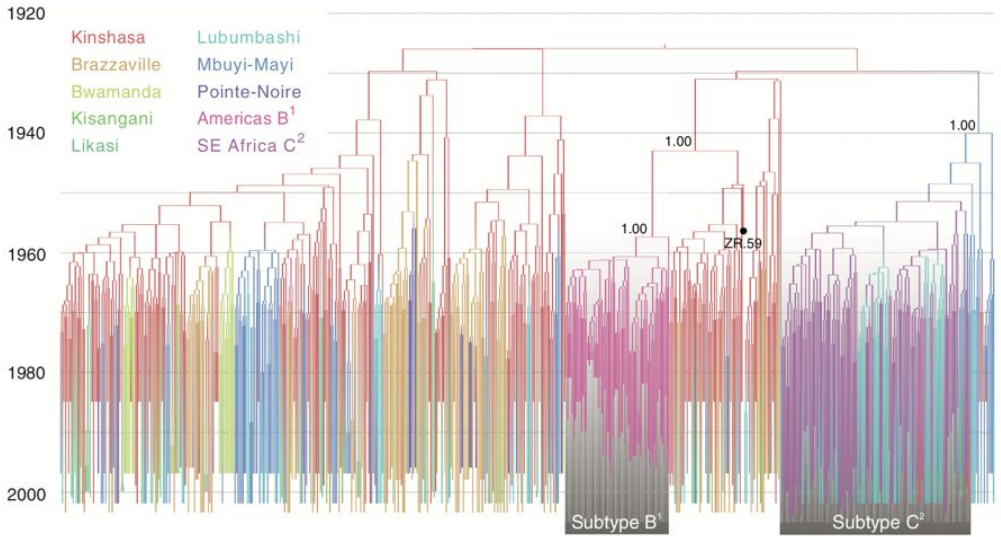

In future lectures, we'll learn much more about how linkage can be used to reconstruct entire transmission networks.

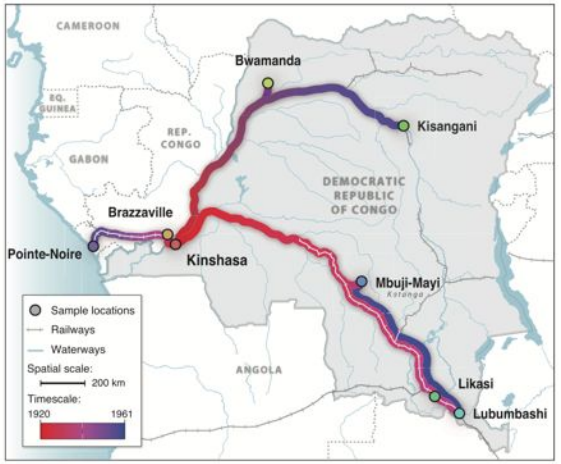
# Phylogenies can tell us about:

- Linkage and the sources of transmission
- The origins of epidemics and new strains
- Past epidemic dynamics
- Pathogen fitness and adaptation

# Origins of the HIV-1 epidemic

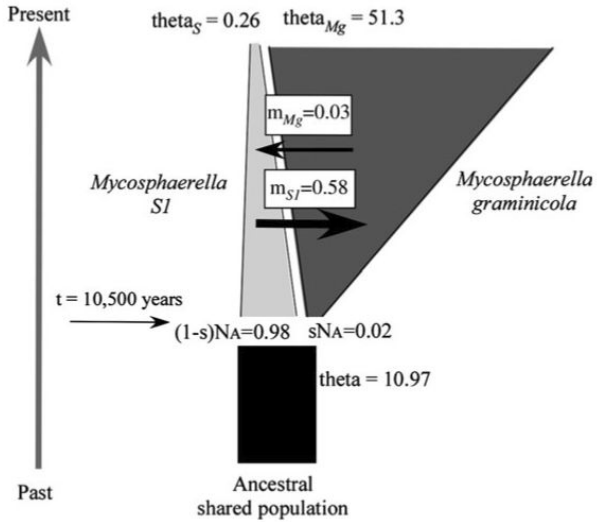
Faria *et al.* (Science, 2014) traced the origins of the HIV-1 epidemic back to the 1920's and 30's in Kinshasa, DRC.



# Origins of *Mycosphaerella graminicola*

Stukenbrock *et al.* (MBE, 2006) traced the fungal pathogen causing septoria leaf blotch on wheat back to 8,000 to 9,000 BC in the Fertile Crescent.




*M. graminicola* on wheat (Wikipedia)

# Neolithic origins of other agro-pathogens

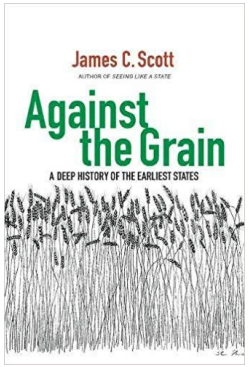
Supports idea that many agriculturally important pathogens arose during the Neolithic transition to farming.

**Table 1** Examples of evolutionary mechanisms by which plant pathogens have emerged in agro-ecosystems over different time scales

| Evolutionary mechanism | Plant pathosystem | Time scale | Reference |
|---|---|---|---|
| **Domestication/host-tracking** | | | |
| | *Mycosphaerella graminicola* on wheat | 10–12,000 years BP | 95 |
| | *Magnaporthe oryzae* on rice | 7000 years BP | 24 |
| | *Phytophthora infestans* on potato | 7000 years BP | 34 |
| | *Ustilago maydis* on maize | 8000 years BP | 72 |
| **Host jump/host shift** | | | |
| | *Magnaporthe oryzae* from Setaria millet to rice | Abrupt evolutionary change, approx. 7000 years BP | 24 |
| | *Rhynchosporium secalis* from wild grasses to barley and rye | Abrupt evolutionary change, approx. 2,000 years BP | 111 |
| | *Phytophthora infestans* from wild *Solanum* species to potato | Abrupt evolutionary change, <500 years BP | 35, 39 |

Stukenbrock and McDonald (Annu. Rev. Phyto., 2008)

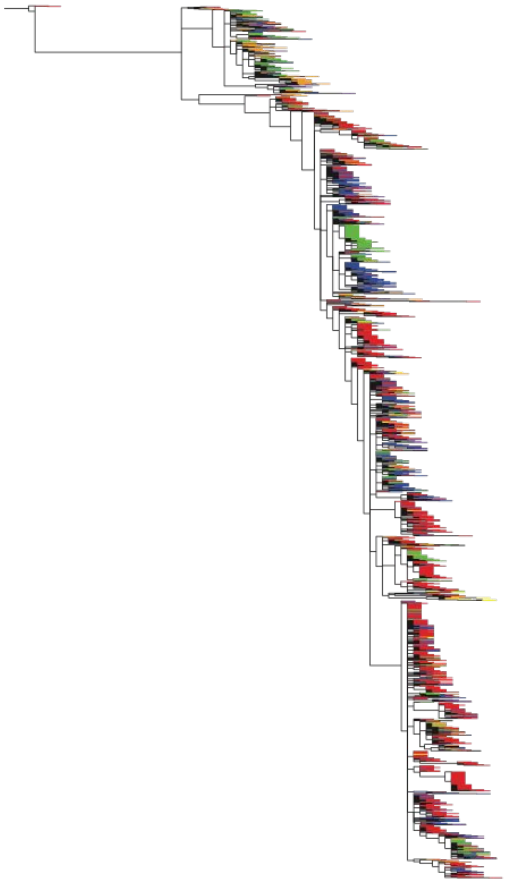
"Neolithic pathogen relocation camps"



# Influenza A (H3N2)

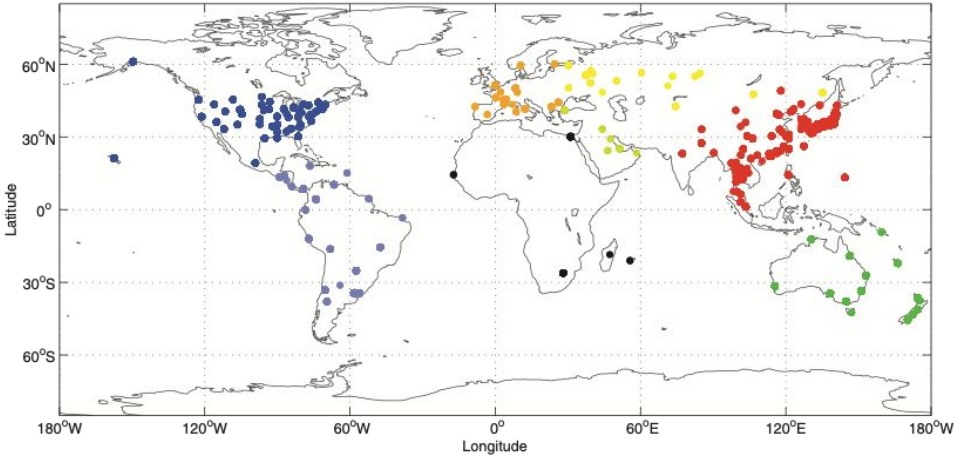
New antigenic variants periodically replace older strains

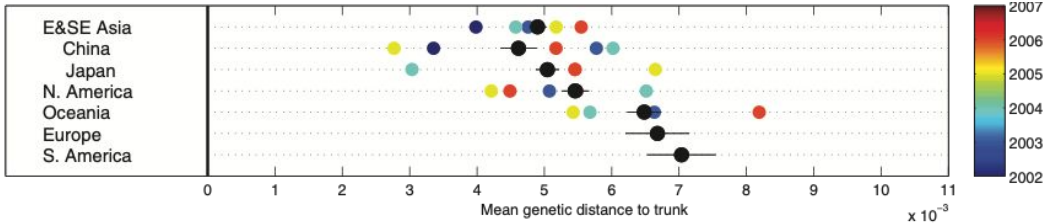
- Antigenic drift leads to a ladder-like structure with a **trunk lineage**
- Drift creates antigenic mismatch with vaccines.
- Where do new antigenic variants arise?

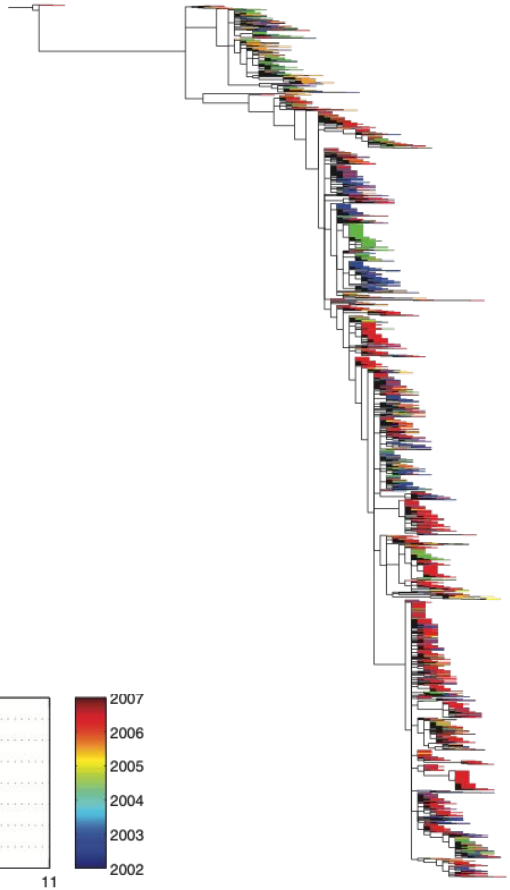


Russell *et al.* (Science, 2008)

# Influenza A (H3N2)

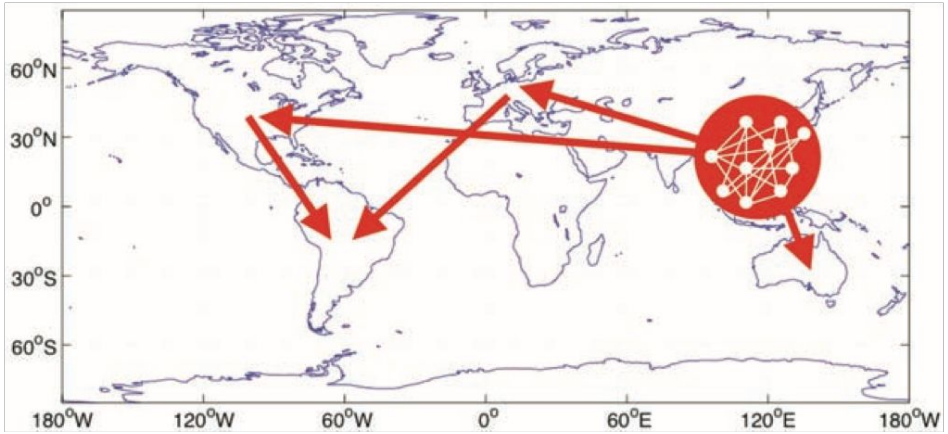


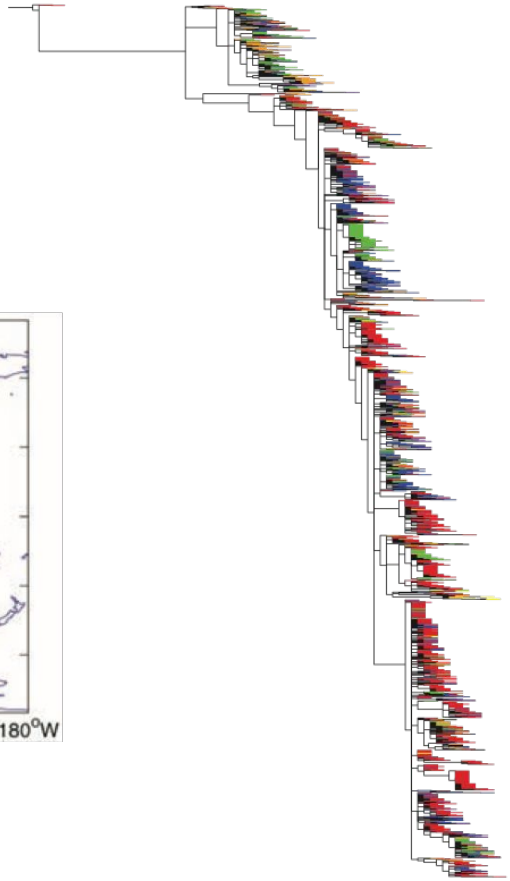



Russell *et al.* (Science, 2008)

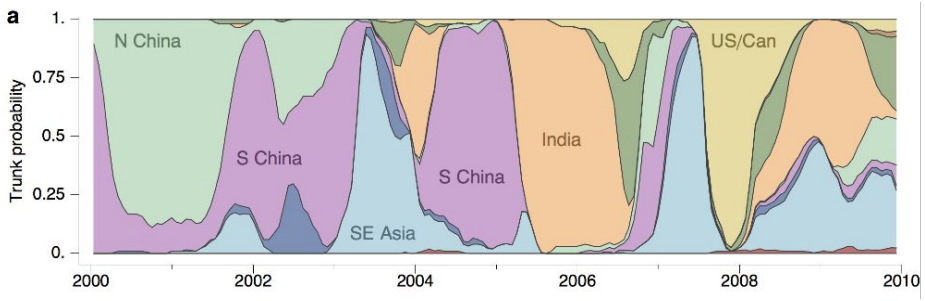
# Influenza A (H3N2)

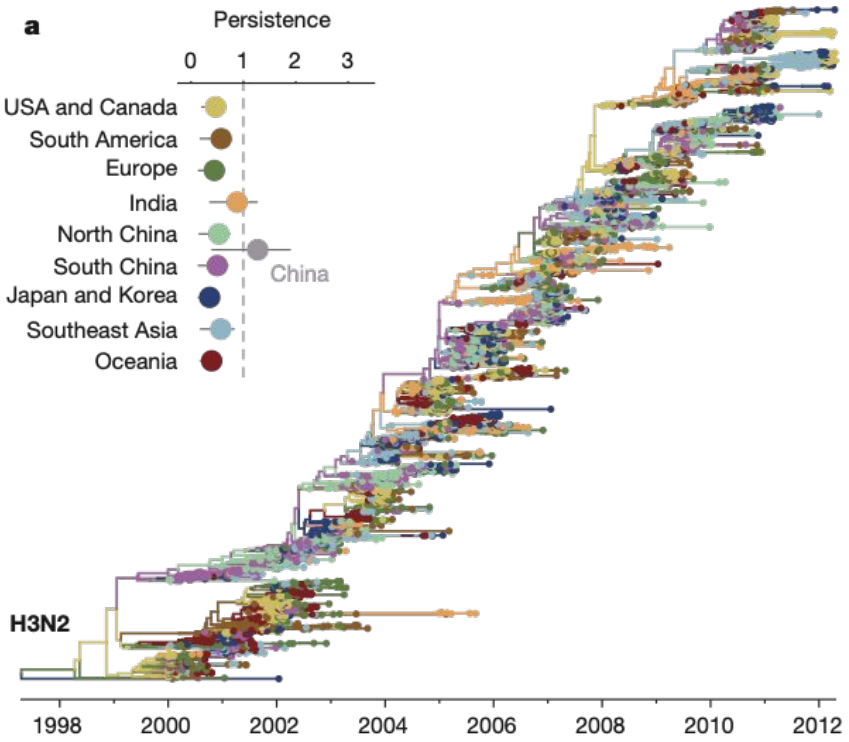


Russell *et al.* (Science, 2008)

# Influenza A (H3N2)



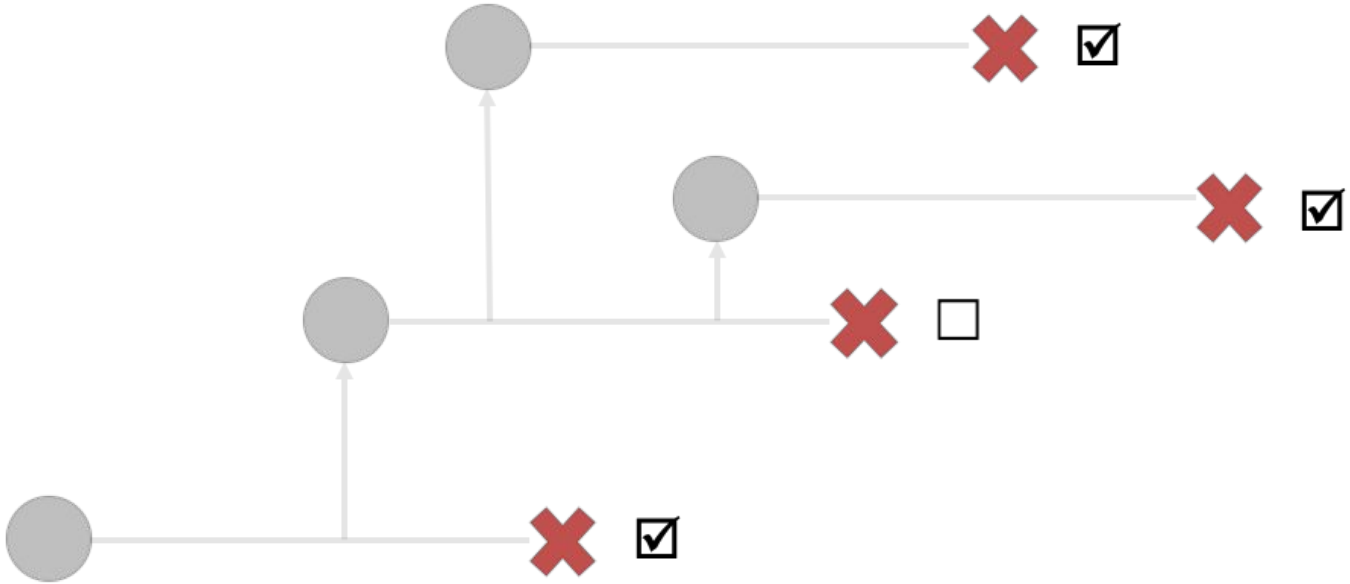




Bedford *et al.* (Nature, 2015)

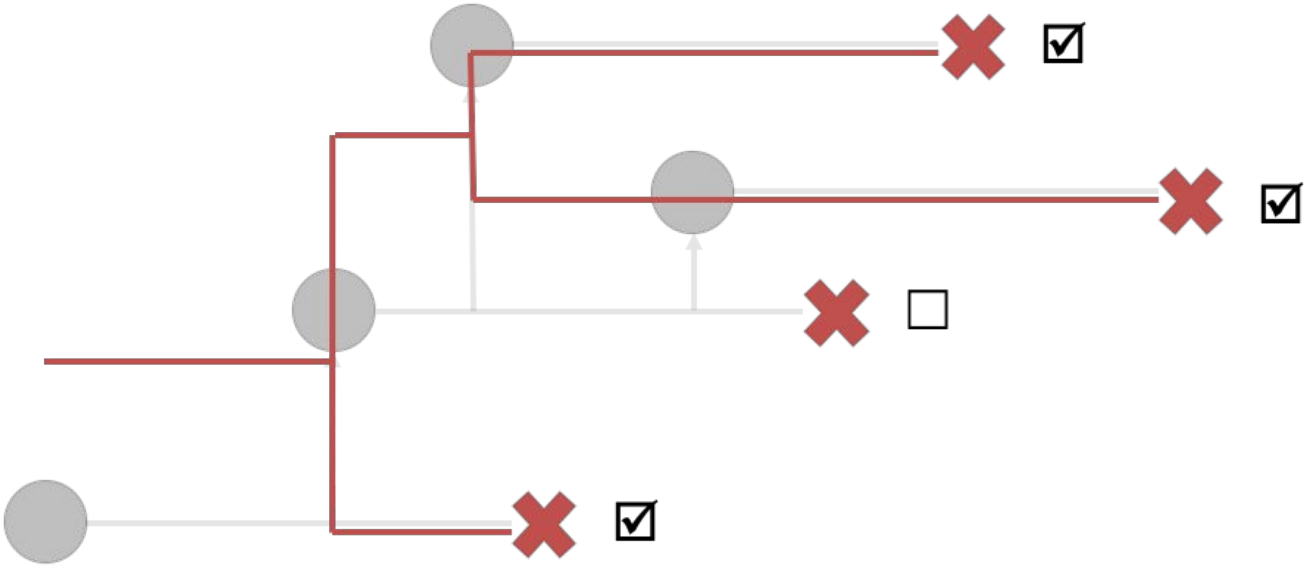
# Phylogenies can tell us about:

- Linkage and the sources of transmission
- The origins of epidemics and new strains
- Past epidemic dynamics
- Pathogen fitness and adaptation

# A simple epidemic example with incomplete sampling



# A simple epidemic example with incomplete sampling



**We only observe transmission events as branching events if we sample both the parent and child lineage descending from the transmission event**

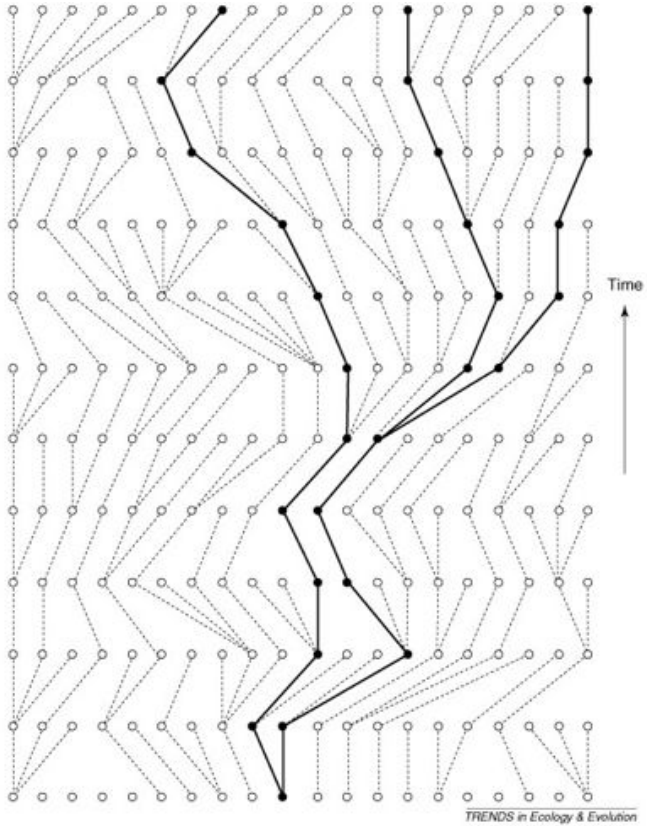
This brings us to phylodynamic modeling

# Phylodynamic modeling in a nutshell

Phylogenies will only contain sampled lineages.

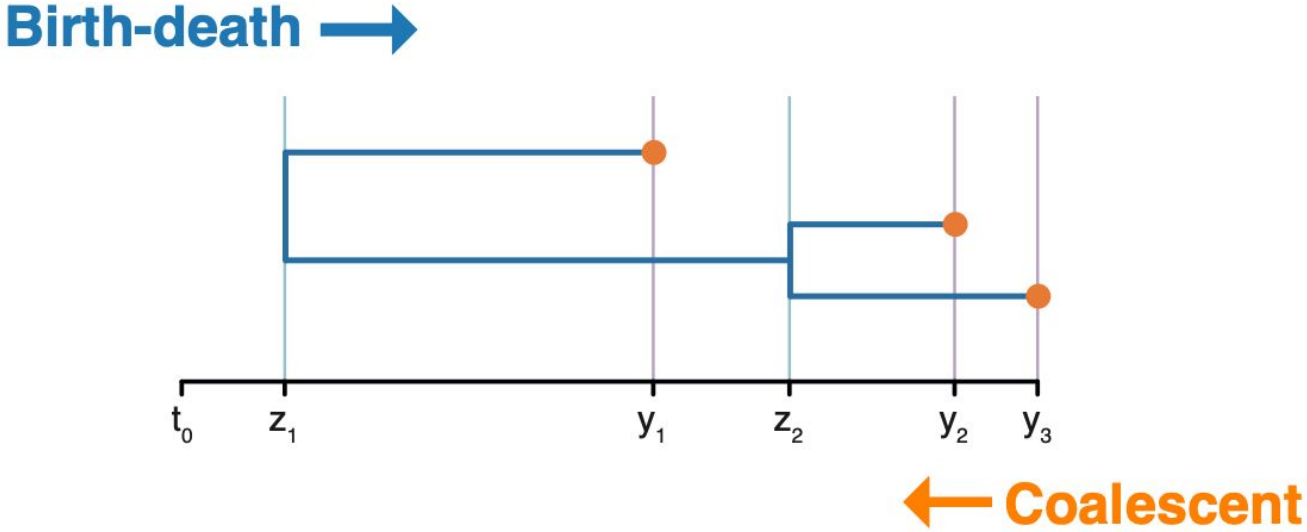
The sampled lineages are embedded within the full ancestral history of the population.

We need a statistical model that allows us to infer the most likely population history from the sampled phylogeny.




Kuhner *et al.* (2008)

# Two types of phylodynamic models



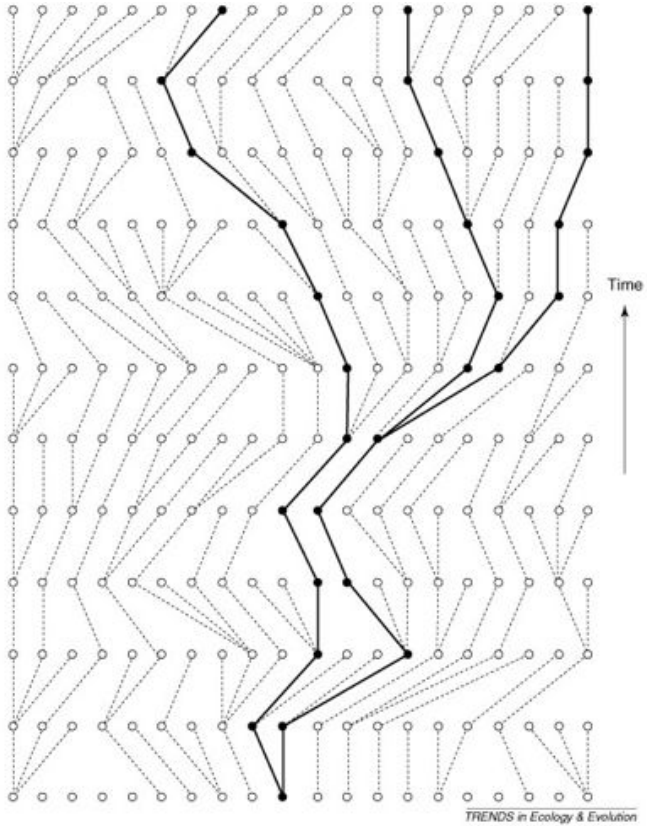
# Coalescent theory

The coalescent traces the ancestry of sampled individuals back in time.

Allows us to relate events observed in the tree to the larger history of a population

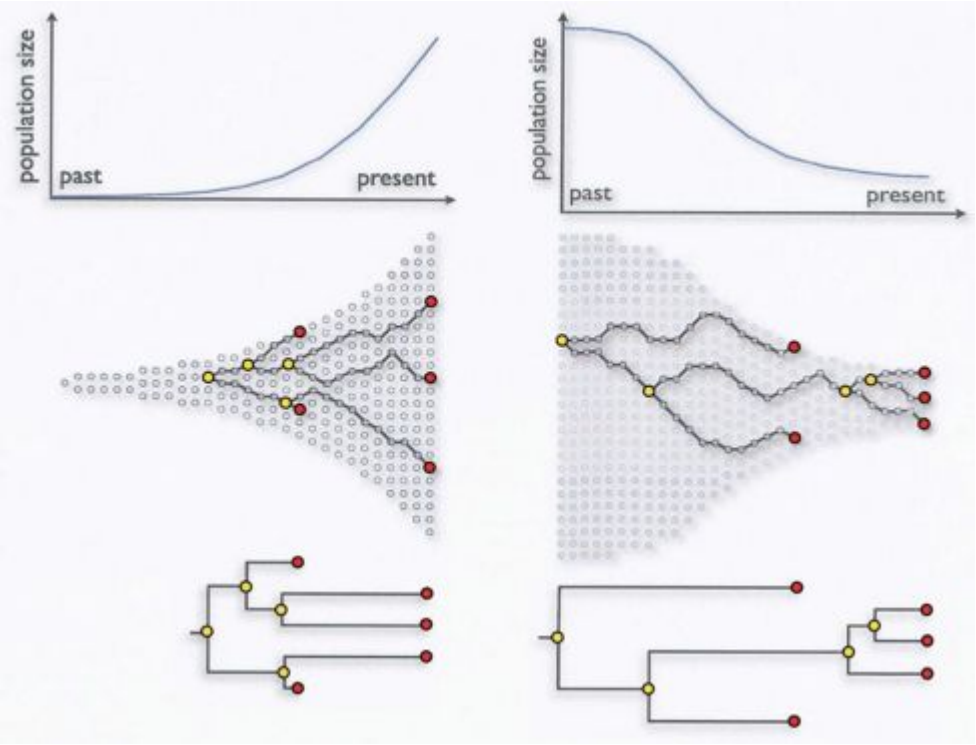
Probability of two lineages coalescing per generation is:

$$p_{coal} = \frac{1}{N}$$




Kuhner *et al.* (2008)

# Reconstructing population dynamics

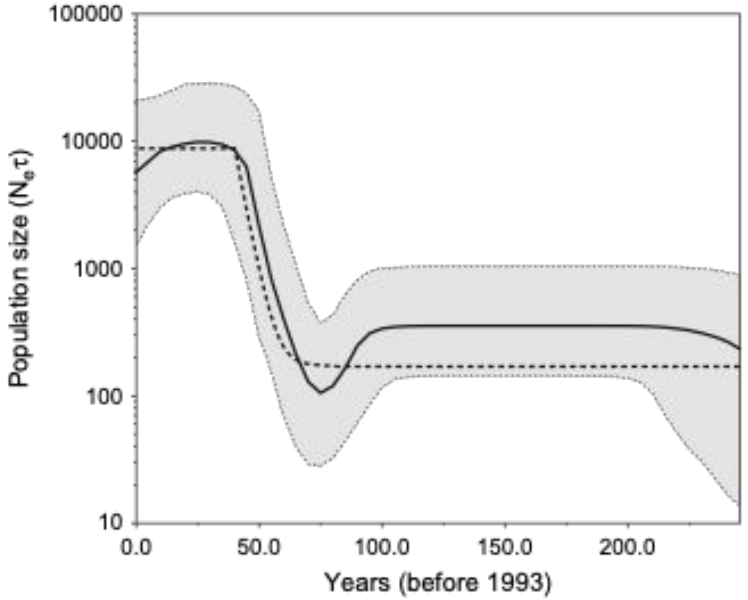


# (Non-)Parametric phylodynamic methods

Nonparametric methods reconstruct effective population sizes ($N_e$) without assuming an explicit population dynamic model

- Bayesian Skyline (Drummond *et al.*, 2005)
- Bayesian Skygrid (Minin *et al.*, 2008)

# Reconstructing dynamics: Hepatitis C



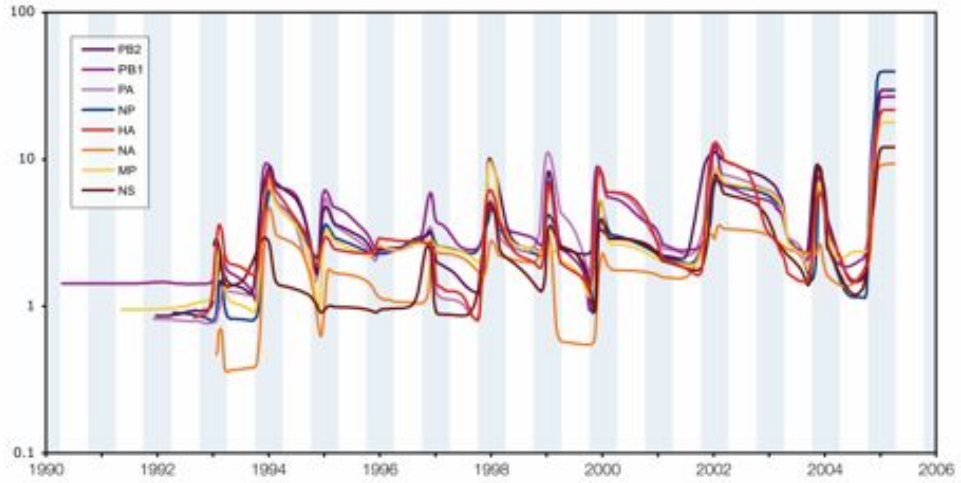

Pybus *et al.* (2003)
Drummond *et al.* (2005)

# Reconstructing dynamics: influenza A

## The genomic and epidemiological dynamics of human influenza A virus

Andrew Rambaut[1], Oliver G. Pybus[2], Martha I. Nelson[3], Cecile Viboud[4], Jeffery K. Taubenberger[3] & Edward C. Holmes[3,4]

**The evolutionary interaction between influenza A virus and the human immune system, manifest as 'antigenic drift' of the viral haemagglutinin, is one of the best described patterns in molecular evolution. However, little is known about the genome-scale evolutionary dynamics of this pathogen. Similarly, how genomic processes relate to global influenza epidemiology, in which the A/H3N2 and A/H1N1 subtypes co-circulate, is poorly understood. Here through an analysis of 1,302 complete viral genomes sampled from temperate populations in both hemispheres, we show that the genomic evolution of influenza A virus is characterized by a complex interplay between frequent reassortment and periodic selective sweeps. The A/H3N2 and A/H1N1 subtypes exhibit different evolutionary dynamics, with diverse lineages circulating in A/H1N1, indicative of weaker antigenic drift. These results suggest a sink–source model of viral ecology in which new lineages are seeded from a persistent influenza reservoir, which we hypothesize to be located in the tropics, to sink populations in temperate regions.**




Rambaut *et al.* (2008)

# (Non-)Parametric phylodynamic methods

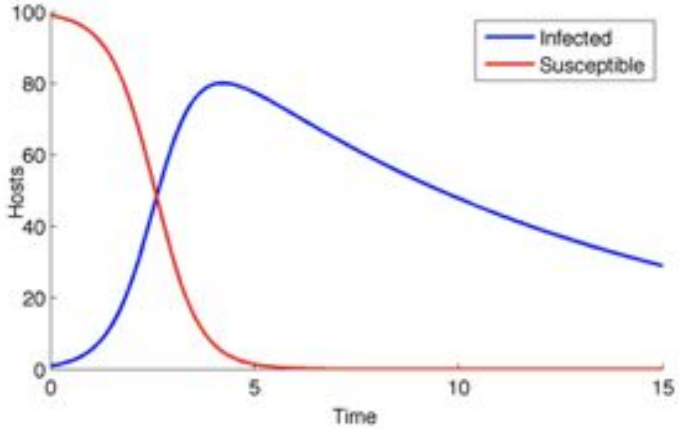
Nonparametric methods reconstruct effective population sizes ($N_e$) without assuming an explicit population dynamic model

- Bayesian Skyline (Drummond *et al.*, 2005)
- Bayesian Skygrid (Minin *et al.*, 2008)

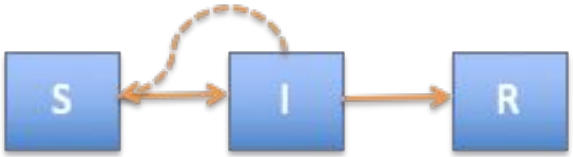
**Alternatively**, we can derive an appropriate phylodynamic model from a mechanistic population dynamic model with relevant epidemiological parameters

# Epidemiological models

# Epidemiological models

# Epidemiological models




$$\frac{dS}{dt} = -\beta \frac{S}{N} I$$
$$\frac{dI}{dt} = \beta \frac{S}{N} I - \nu I$$
$$\frac{dR}{dt} = \nu I$$

# SIR-type coalescent model

The total rate at which transmission/coalescent events occur in the population depends on the incidence of new infections:

$$f(t) = \beta S(t) I(t)$$

Given a transmission event occurs, the probability that two particular lineages coalesce is:

$$p_{coal} = \frac{2}{Y(t)^2} = \frac{2}{I(t)^2}$$

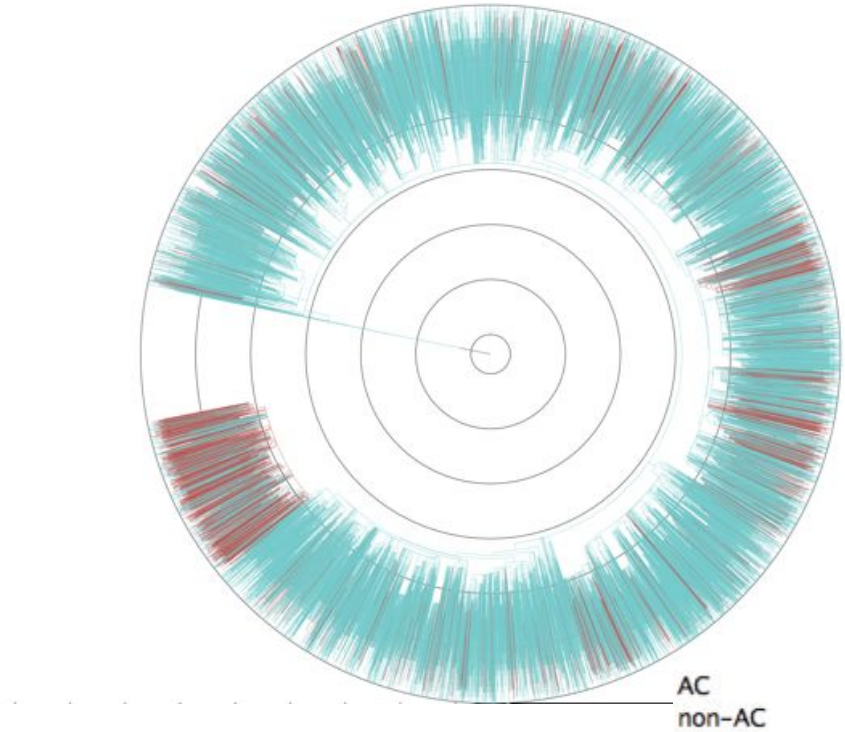
The pairwise coalescent rate is therefore:

$$\lambda(t) = \frac{2f(t)}{Y(t)^2} = \frac{2\beta S(t) I(t)}{I(t)^2} = \frac{2\beta S(t)}{I(t)}$$

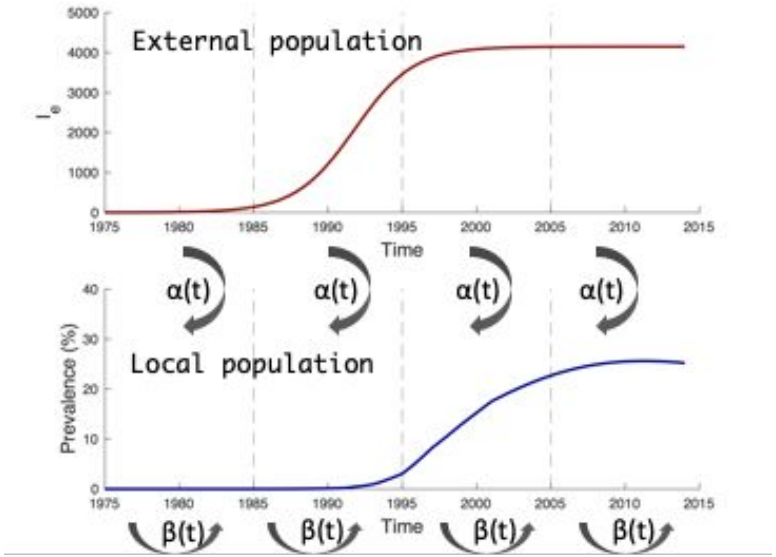
Volz *et al.* (Genetics, 2009)
Volz (Genetics, 2012)

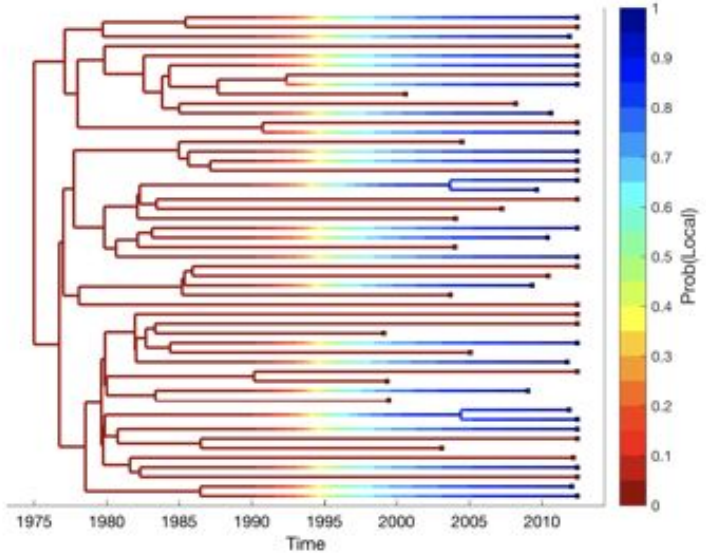
We can formulate epidemic models that we can then fit to phylogenies to estimate parameters of interest.

# HIV in rural Kwa-Zulu Natal

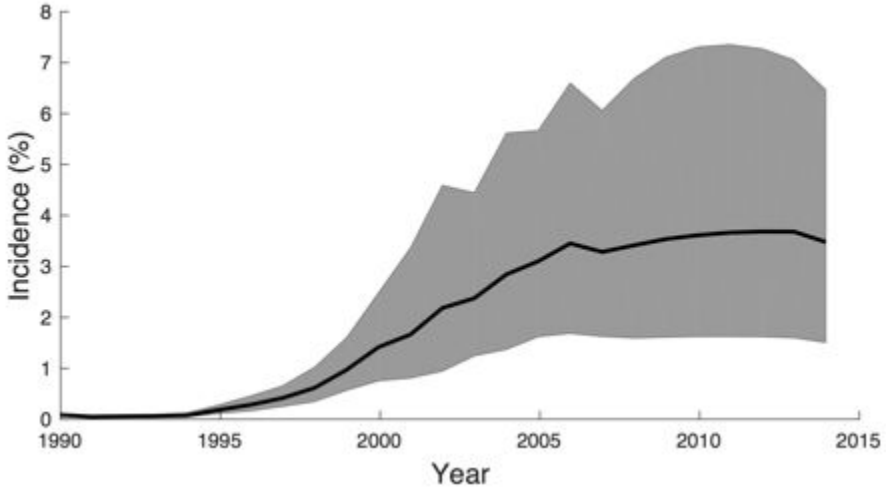



# A simple two-patch SIR model for HIV



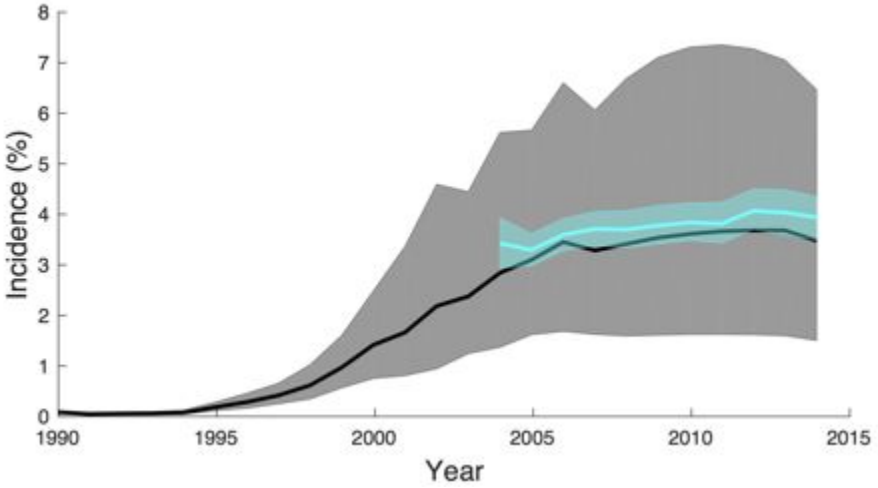
# Phylodynamic estimates of HIV incidence




Rasmussen et al. (Virus Evolution, 2018)

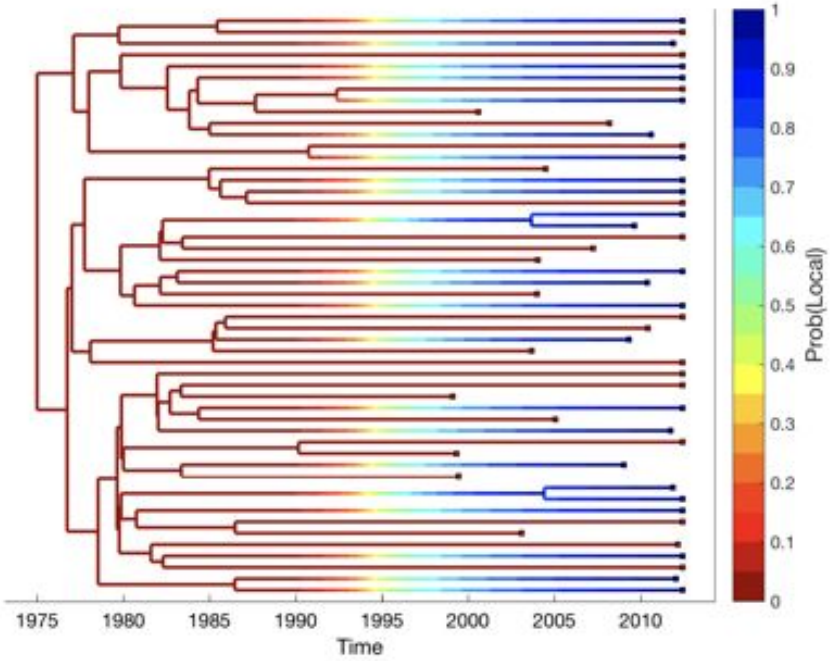
# Phylodynamic estimates of HIV incidence

Inferred incidence of 3-4% per year almost perfectly coincides with population-based surveillance data.

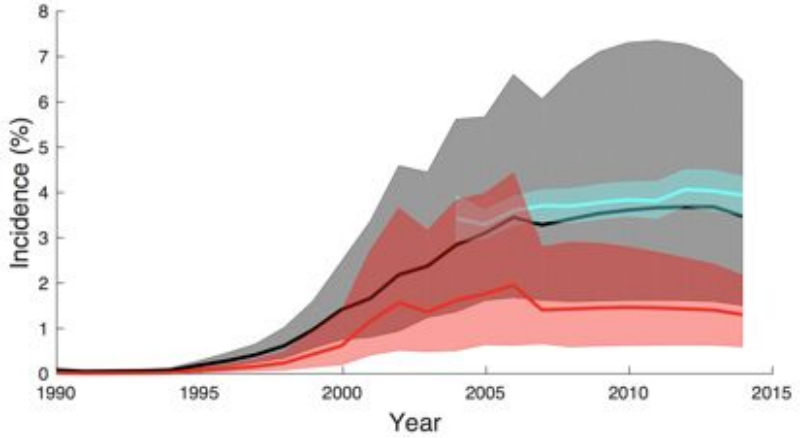



Rasmussen et al. (Virus Evolution, 2018)

# Tracking lineage movement

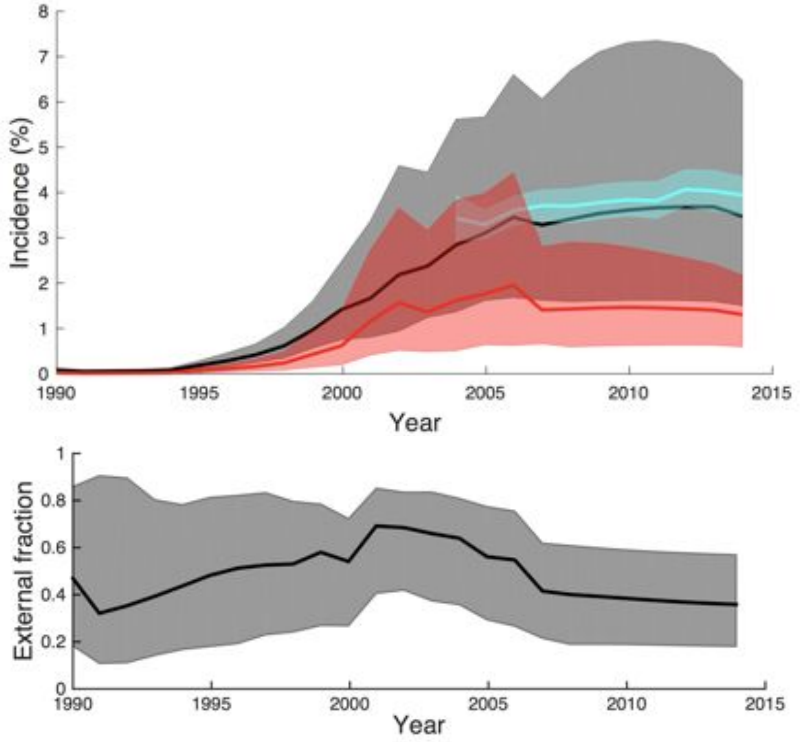


# Incidence due to external introductions




Rasmussen et al. (Virus Evolution, 2018)

# Incidence due to external introductions




As of 2014, 35% of new infections were attributed to external introductions.

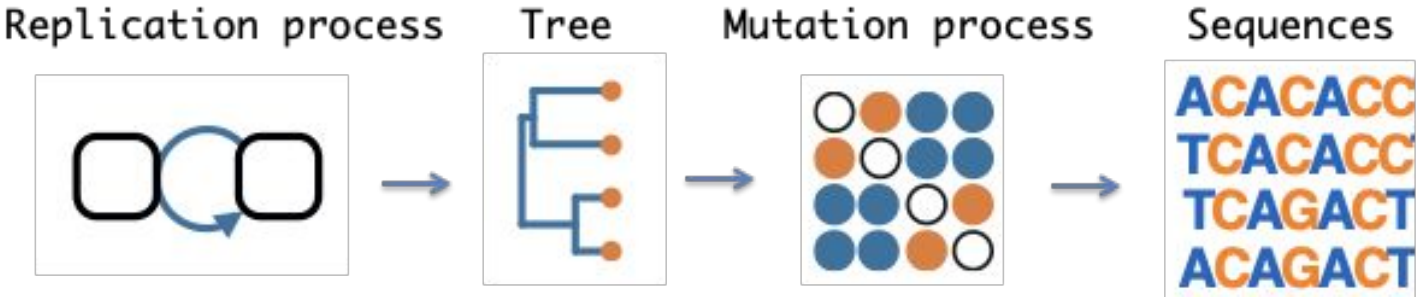
Rasmussen et al. (Virus Evolution, 2018)

# Phylogenies can tell us about:

- Linkage and the sources of transmission
- The origins of epidemics and new strains
- Past epidemic dynamics
- Pathogen fitness and adaptation

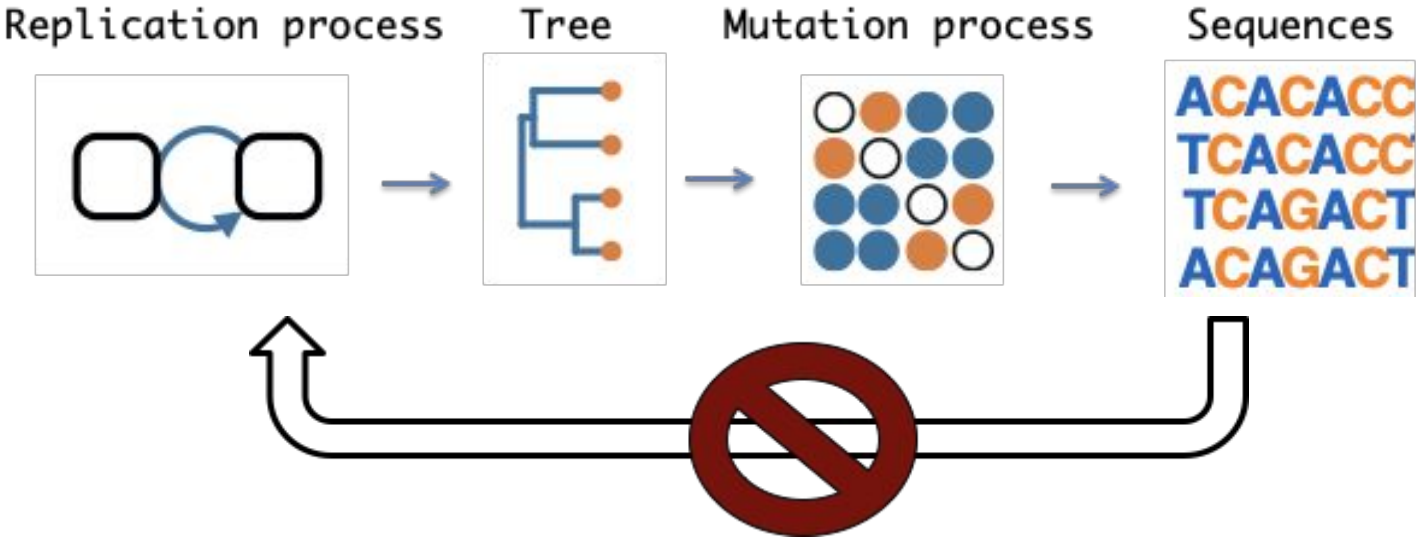
# The big assumption of phylogenetic models

Most phylogenetic models ignore non-neutral evolution:



# The big assumption of phylogenetic models

Most phylogenetic models ignore non-neutral evolution because they don't allow mutations to feedback and affect the tree generating (replication) process.



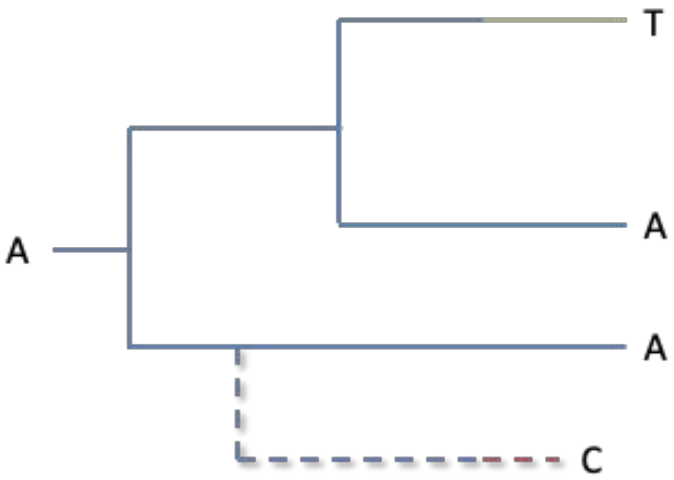
# Phylodynamics with selection

Selection for better adapted strains strongly shapes the phylogenetic history of many different pathogens.

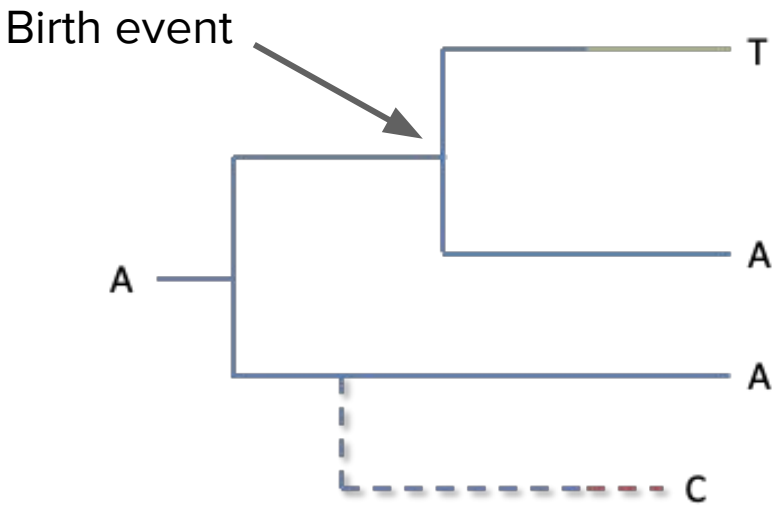

Grenfell *et al.* (Science, 2004)

# Multi-type birth-death models

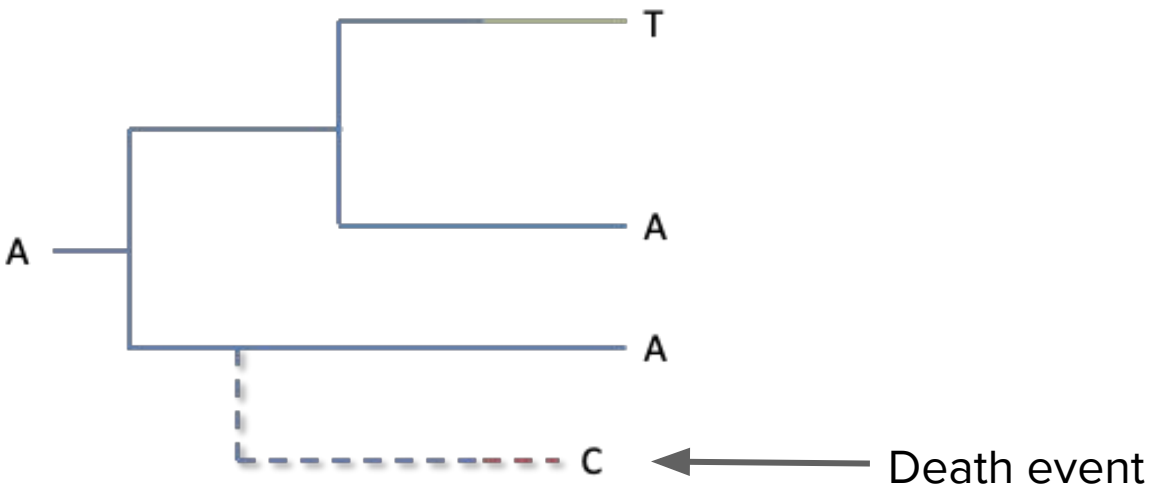
Provide one way of incorporating adaptive (non-neutral) evolution into phylogenetic models.



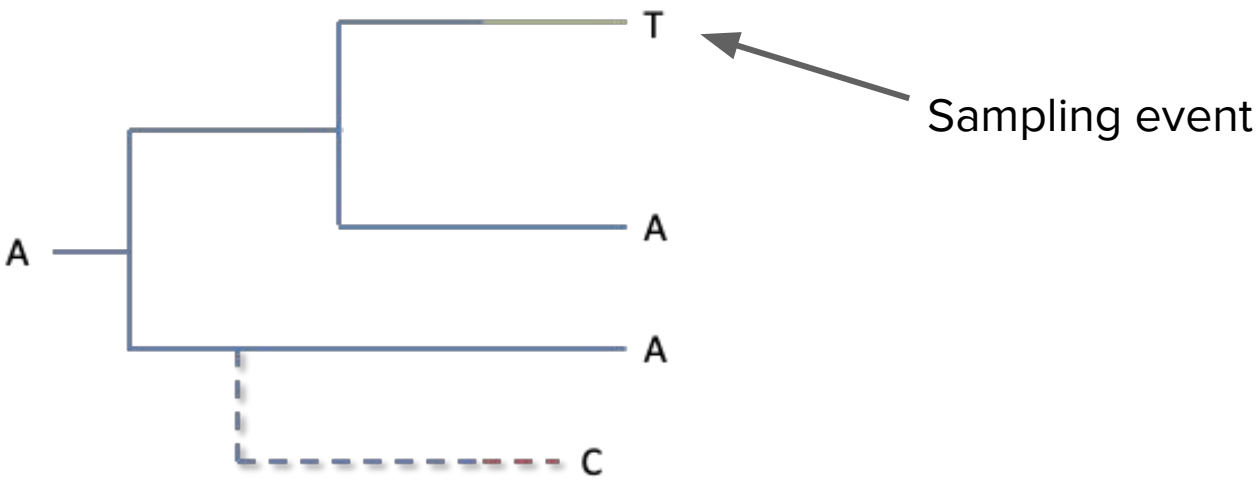
# Multi-type birth-death models



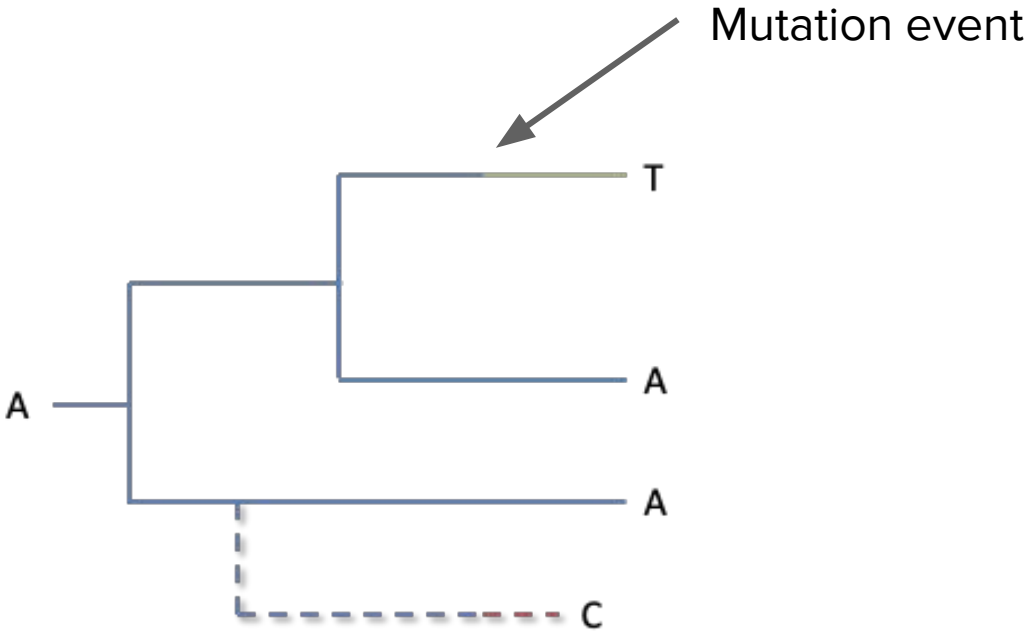
# Multi-type birth-death models



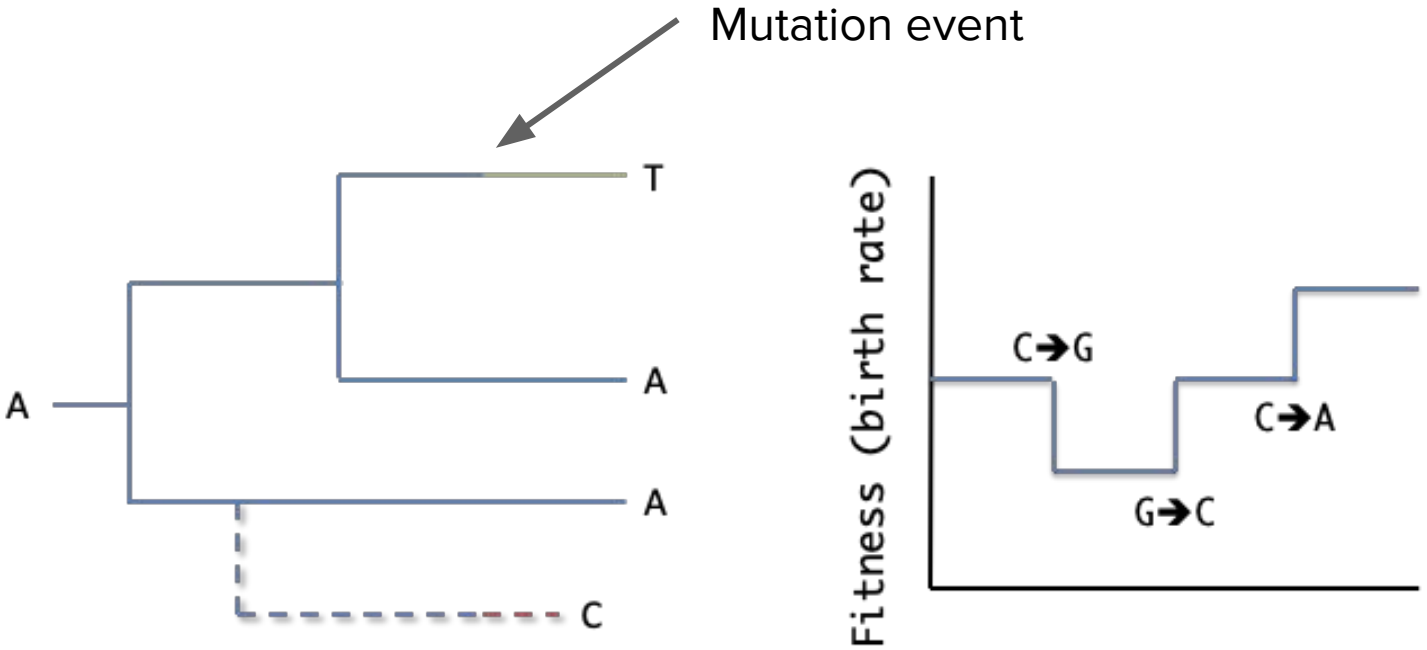
# Multi-type birth-death models



# Multi-type birth-death models

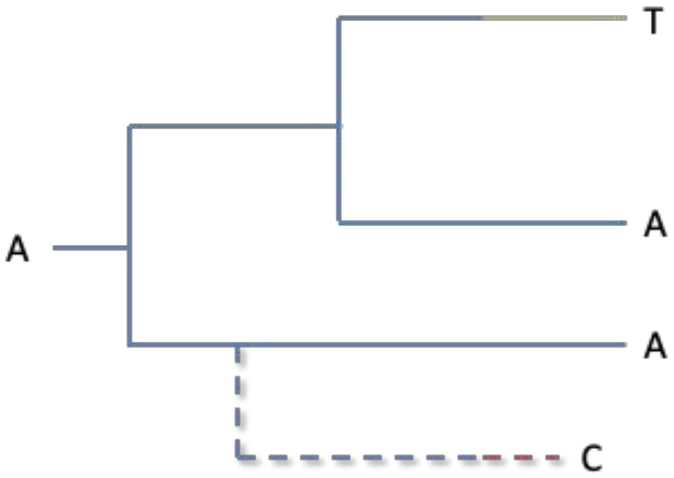


# Multi-type birth-death models



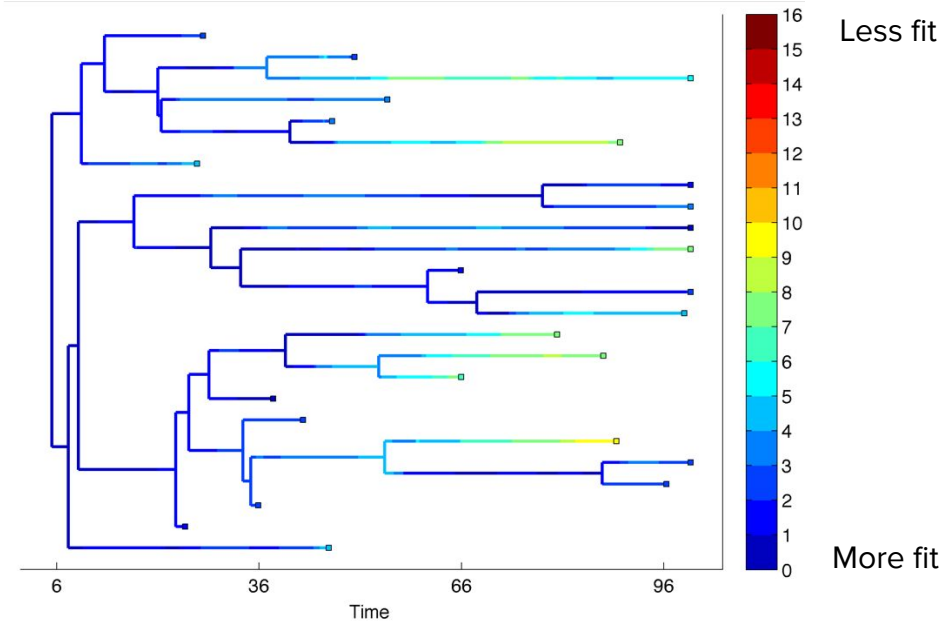
# Multi-type birth-death models

MTBD models allow us to compute the **joint likelihood** that both the tree and the observed tip genotypes evolved exactly as observed (Stadler and Bonhoeffer, 2013).



# Fitness shapes trees

More fit lineages will branch (reproduce) more and leave behind more sampled descendants than less fit lineages.



# Fitness of HIV drug resistance mutations

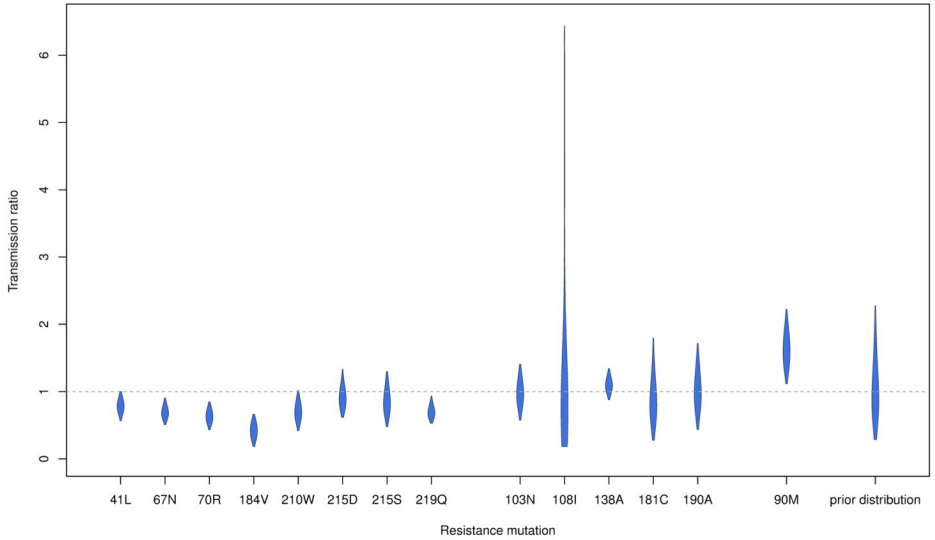
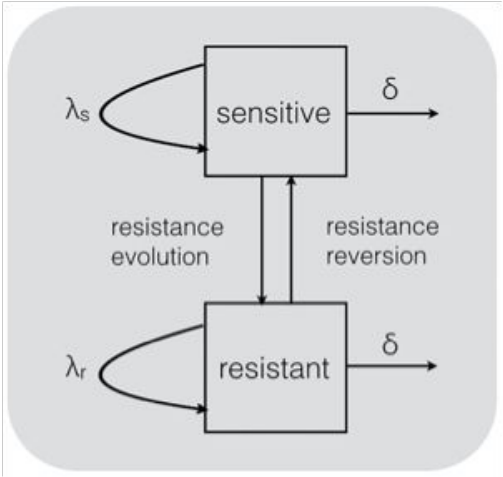



**Table 1. Resistance mutations with numbers of corresponding clusters and samples, related drugs and drug usage dates within Switzerland.**

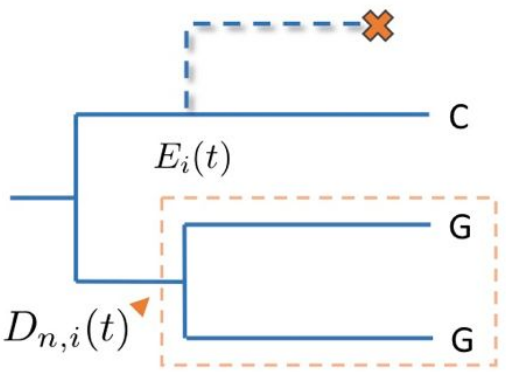
| | nRTI | | | | | | | | | NNRTI | | | | | PI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Resistance mutation | 41L | 67N | 70R | 184V | 210W | 215D | 215S | 215Y | 219Q | 103N | 108I | 138A | 181C | 190A | 90M |
| Number (#) of clusters of size $\geq 2$ | 56 | 23 | 19 | 35 | 18 | 18 | 16 | 25 | 20 | 25 | 10 | 46 | 8 | 8 | 14 |
| # Sequences in clusters | 927 | 667 | 712 | 1011 | 481 | 569 | 494 | 807 | 605 | 725 | 334 | 1014 | 329 | 311 | 389 |
| # Resistant samples in clusters | 93 | 39 | 26 | 44 | 26 | 41 | 31 | 28 | 28 | 38 | 11 | 109 | 10 | 12 | 38 |
| Drug (SHCS drug codes) | AZT D4T | AZT D4T | AZT D4T | 3TC ABC FTC | AZT D4T | AZT D4T | AZT D4T | AZT D4T | AZT D4T | NVP EFV | NVP EFV | RPV | NVP EFV ETV RPV | NVP EFV | NFV SQV |
| Drug usage $\geq 1\%$ | 1987 | 1987 | 1987 | 1995.5 | 1987 | 1987 | 1987 | 1987 | 1987 | 1997 | 1997 | 2013 | 1997 | 1997 | 1996 |
| Drug usage $< 1\%$ | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 2008 |



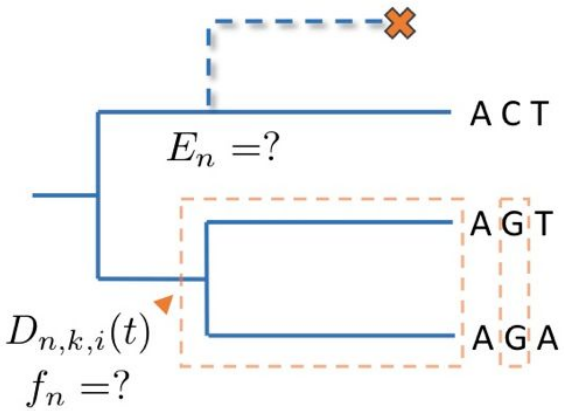

Kühnert *et al.* (PLoS Pathogens, 2018)

# Marginal fitness birth-death models

The Multi-Type Birth-Death Model



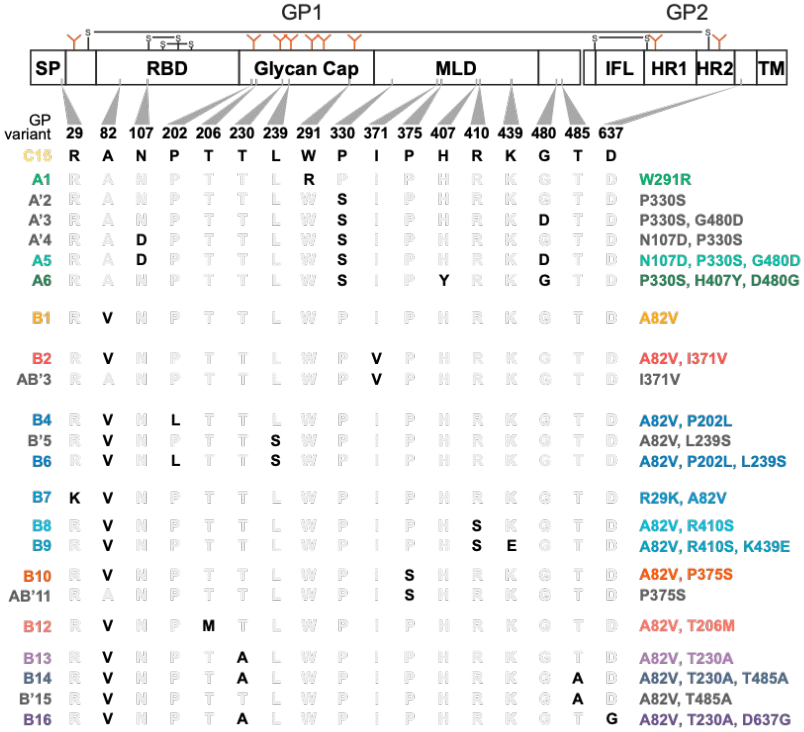

The Marginal Fitness Birth-Death Model

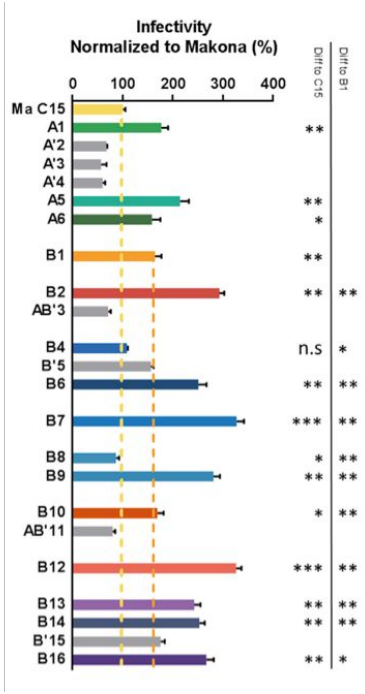



1. $\hat{\omega}_{n,\mathrm{ACT}} = \omega_{n,1,\mathrm{A}} \times \omega_{n,2,\mathrm{C}} \times \omega_{n,3,\mathrm{T}}$
2. $f_n \approx \sum_{g \in \mathcal{G}} f_g \hat{\omega}_{n,g}$
3. $E_n \approx E_u$

Rasmussen and Stadler (ELife, 2019)

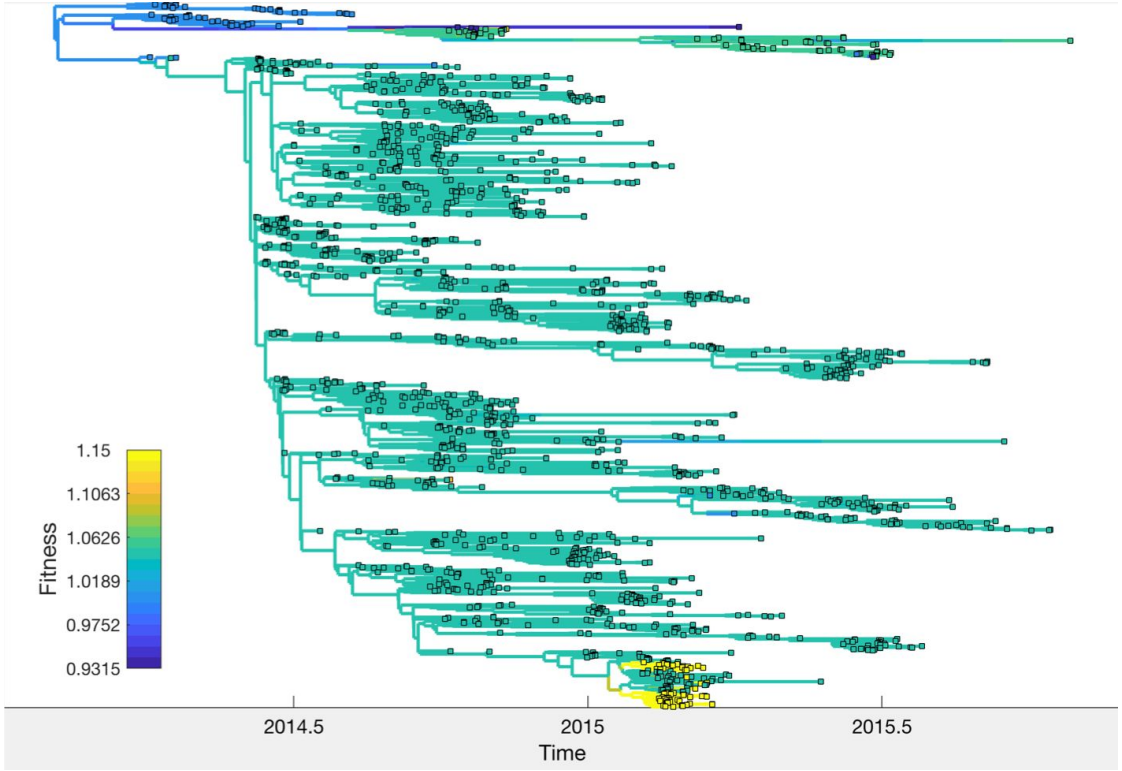
# Adaptation of Ebola virus to humans




Urbanowicz *et al.* (Cell, 2016)

# Adaptation of Ebola virus to humans




Rasmussen and Stadler (ELife, 2019)

# Phylogenies can tell us about:

- Linkage and the sources of transmission
- The origins of epidemics and new strains
- Past epidemic dynamics
- Pathogen fitness and adaptation

# My biases

Tend towards directly transmitted RNA viruses

What about other microbial pathogens?

- Prokaryotic and fungal pathogens
- Animal and plant pathosystems
- Alternative transmission pathways

# For Wednesday

Bring your laptops!

Try to install RAxML ahead of time

If you’re interested in doing Python exercises, install Python (with Anaconda) and Biopython.