An introduction to phylodynamics

Matthew Hall

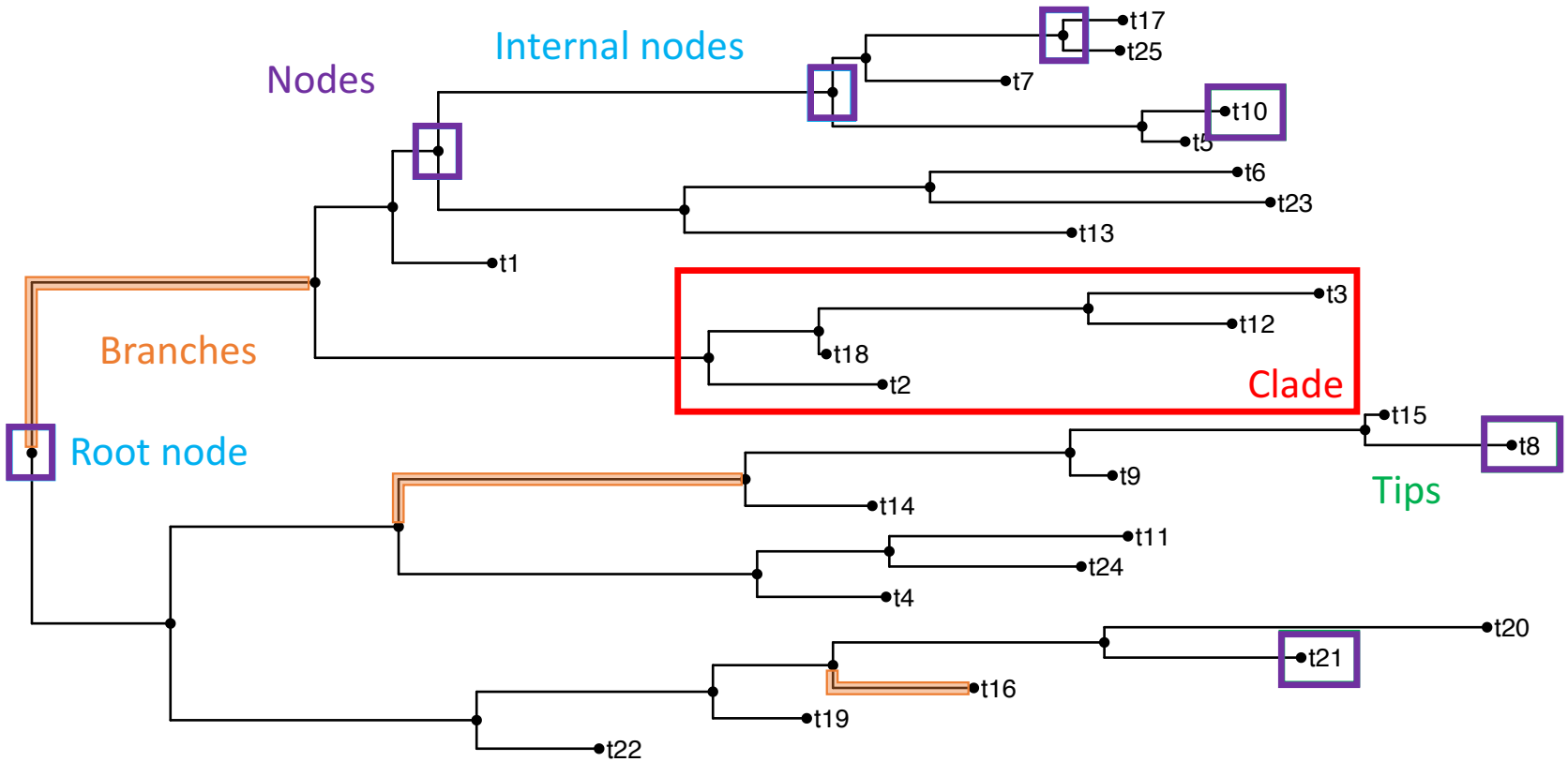
PANGEA webinar series

19/2/19

Recap on phylogenetics

- Phylogenetics is the study of evolutionary relationships between organisms
  - In our case, the organisms are pathogens
- The principle is that the greater the similarity between organisms, the more closely related they are (and hence the more recently they shared a common ancestor)
- Genetic data is used to construct *phylogenetic trees*, depicting the relationships between the ancestors of our samples
  - Usually DNA or RNA sequence data these days
- Next-generation sequencing technology is making the acquisition of molecular data much faster and cheaper than it was in the past

Example phylogeny

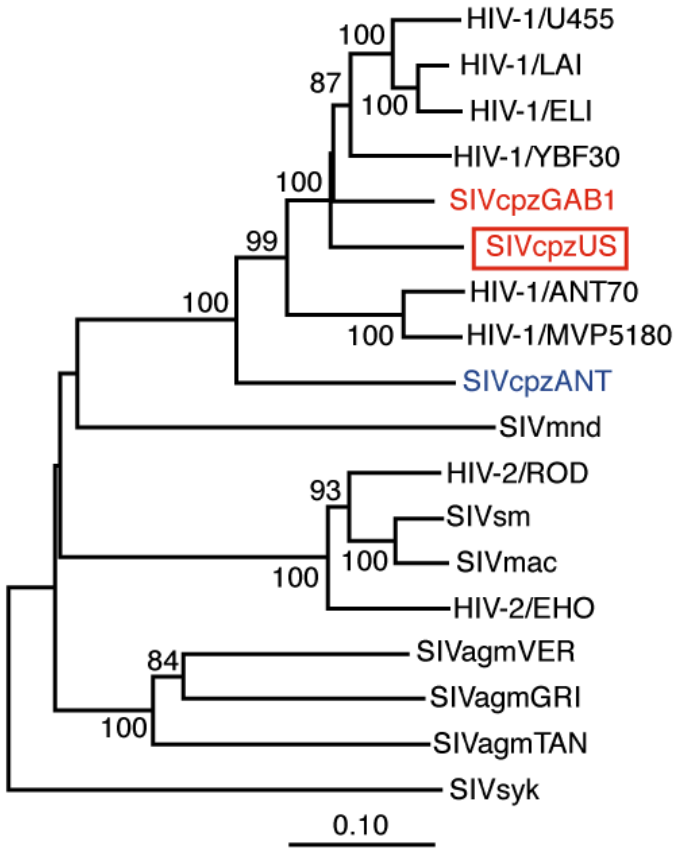


Recap on phylogenetics

- The branch lengths of many phylogenies represent “genetic distance”, a measure of the amount of mutation that have happened
- Many methods and packages are used to create these:
  - Neighbour-joining
  - Maximum parsimony
  - Maximum likelihood
    - Probably the most common
    - PhyML, RAxML and IQ-TREE are popular packages
    - FastTree is a very fast *approximate* ML method
  - Bayesian
    - MrBayes, ExaBayes, RevBayes

Phylogenetic uncertainty

- You can never know your phylogeny is “right” without observing the ancestry
- Sometimes many ancestries are about equally likely
- Phylogenetic uncertainty is handled by:
  - Bootstrapping (non-Bayesian methods)
  - Summarising the posterior distribution (Bayesian methods)
- In either case, the analysis produces many trees and a summary tree is annotated by how many of that set agree with it
- Phylogenetic analysis restricted to a single tree is not ideal (but sometimes the only choice)



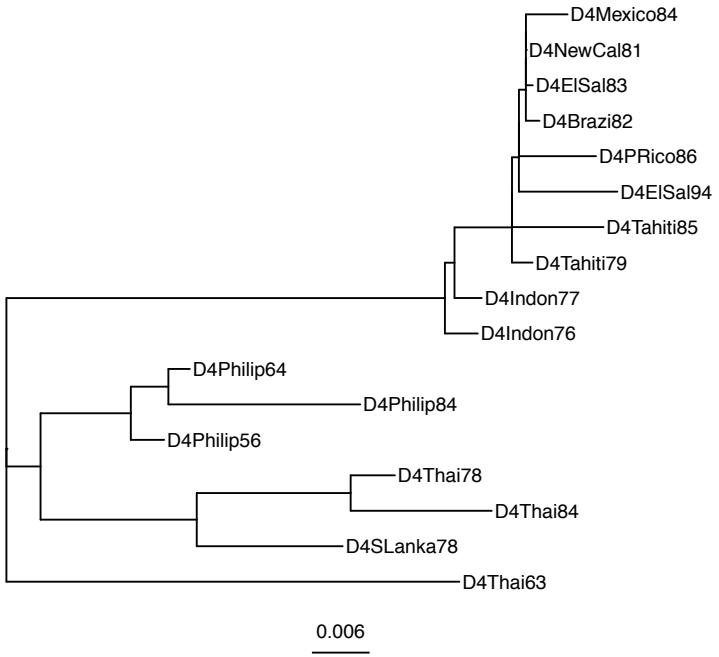
(Gao et al., *Nature*, 1999)

Dating phylogenetic events

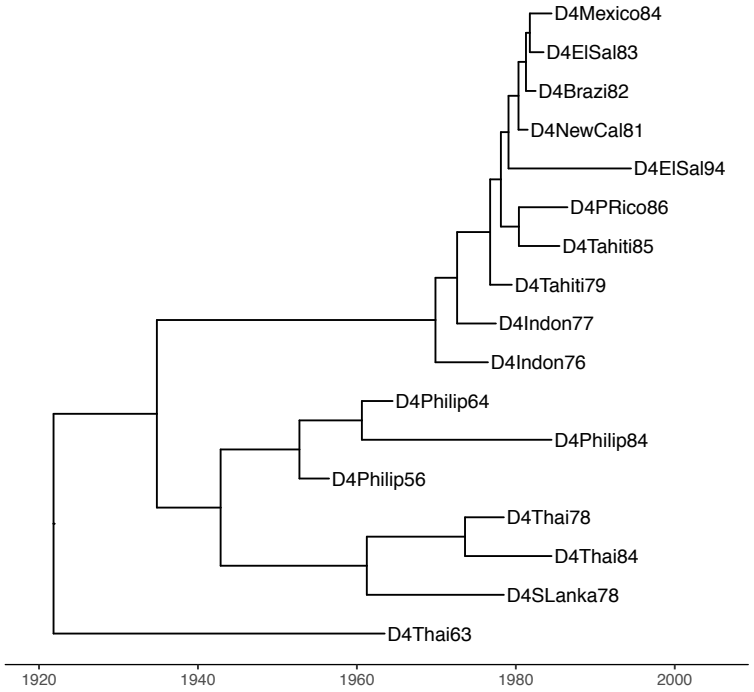
- The phylogeny reconstructs the ancestry of the samples, but on its own it cannot date the events indicated
- If we can date internal nodes, we can estimate when common ancestors existed and when transmissions occurred
- If all our samples were collected over a short period then we require some other information to perform dating
  - Macroorganisms usually breed and mutate too slowly for measurable changes to be observable over a useful timescale
    - Ancient DNA or other archaeological/paleontological findings (for example) have to be used instead
- But pathogen lifespans are short and many (especially RNA viruses) mutate fast and populations observably evolve over short timescales
    - We can use the difference in sampling dates to estimate the rate at which mutations occur

Dating phylogenetic events

Undated



Dated



Molecular clocks

- Mutation is not a deterministic process
  - Longer phylogeny branches do not automatically put a sample further forwards in time
- Instead, mutation is assumed to occur at a rate per unit of time, according to a *molecular clock*
- The simplest version of the molecular clock is the *strict clock* which assumes that this rate is constant over time for a given sample
- This is usually an oversimplification, so various forms of *relaxed clock* are available that allow rates to vary in different regions of the phylogeny
- Not all datasets display behaviour consistent with a molecular clock at all
  - For example, recombination, convergent evolution or simply lack of sufficient variation can remove the signal
  - The presence of a molecular clock signal can be investigated with e.g. TempEst (Rambaut et al., *Virus Evol*, 2018)

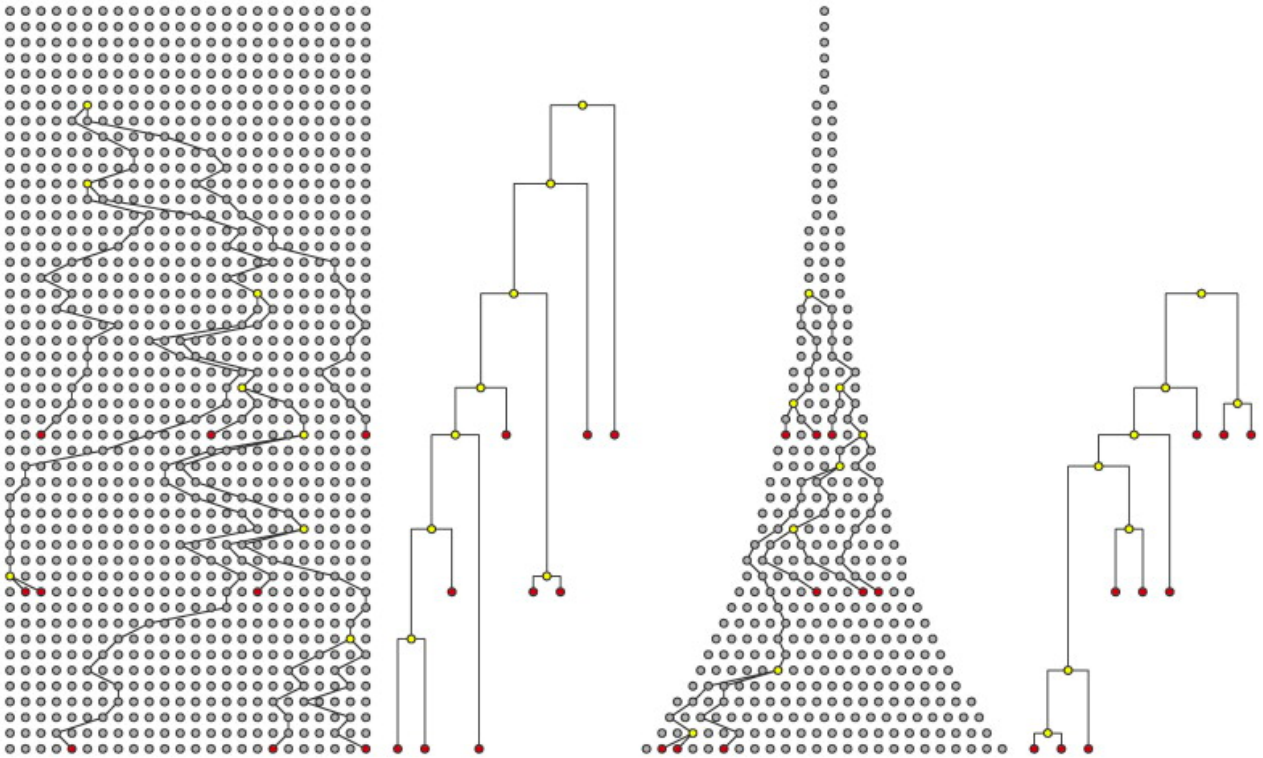
Enter phylodynamics

- The term was coined by Grenfell et al., *Science*, 2004
- The “melding of immunodynamics, epidemiology, and evolutionary biology”
  - The “immunodynamics” bit is for another time
- A dated phylogeny is a (partial) history of a set of pathogen lineages
  - Tips represent samples
  - Internal nodes represent common ancestors of the tips
    - These is often assumed to also represent transmissions between two hosts (which need not be sampled hosts, just ancestors of sampled hosts in the transmission chain)
  - Sample dates are used to calibrate the timings of those ancestors or transmissions according to a molecular clock
- If we have a mathematical model of the process that generates the tree, we can then use sequence data to estimate the parameters of that model and learn about pathogen dynamics

What kinds of models are assumed to generate the trees?

- Broadly, three classes
  1. Population-genetic coalescent models
  2. Forwards-time epidemiological models
  3. Epidemiological coalescent models
- Other related topics
  1. Phylogeography
  2. Transmission tree reconstruction

Population-genetic coalescent models



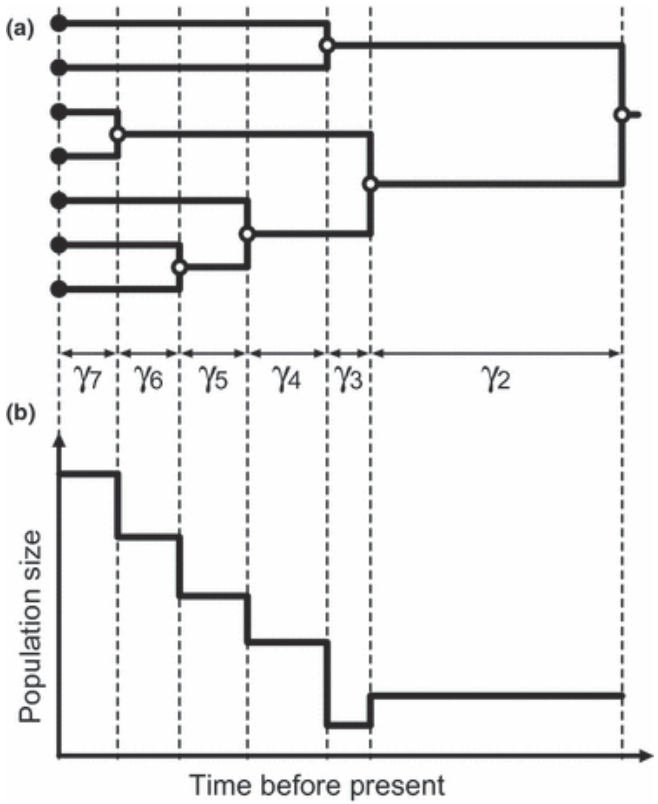
(Kühnert et al., *Inf Genet Evol*, 2011)

- Key principle: in a small population, two individuals are more likely to share an ancestor in the previous generation
- If it takes a long time for two lineages to coalesce, the population must have been large (assuming free mixing)
- We can use the distribution of internal node times (common ancestors) to learn about population size

Population-genetic coalescent models

- In the simplest case, the population size is assumed to be constant and that size (or "effective" size) is estimated
- Alternatively, we can assume it obeys a parametric function (e.g. exponential growth or logistic growth)
- Best of all, we can divide the timeline and estimate sizes separately in each period
- Skyline and associated models (skyride, skygrid)

(Ho and Shapiro, *Mol Ecol Resour*, 2011)



The structured coalescent

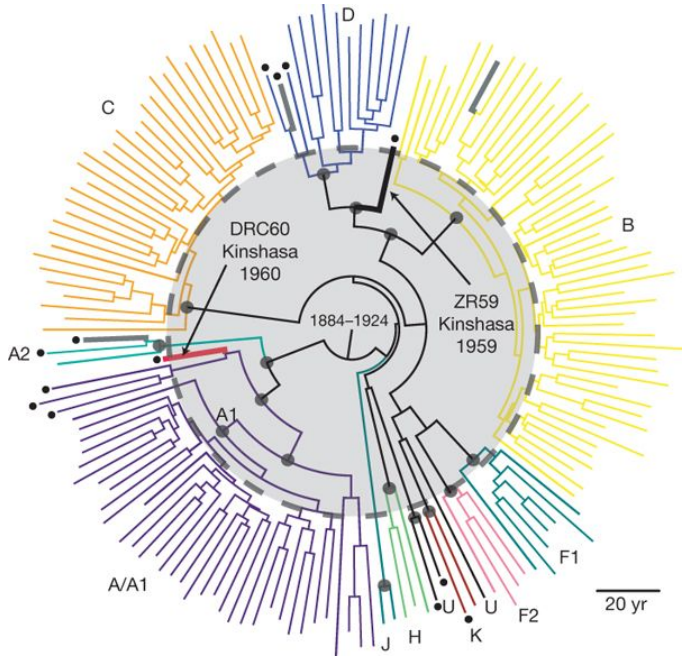
- The standard coalescent assumes that the population is freely-mixing
  - All ancestors are equally likely for any individual in the population
  - This can cause sampling bias issues if some populations are oversampled
- The structured coalescent splits the population into *demes* and allows for ancestry within them and migration between them
  - Individual deme sizes and migration rates may be estimated
  - Skyline models have not yet appeared
  - Works best for small numbers of demes

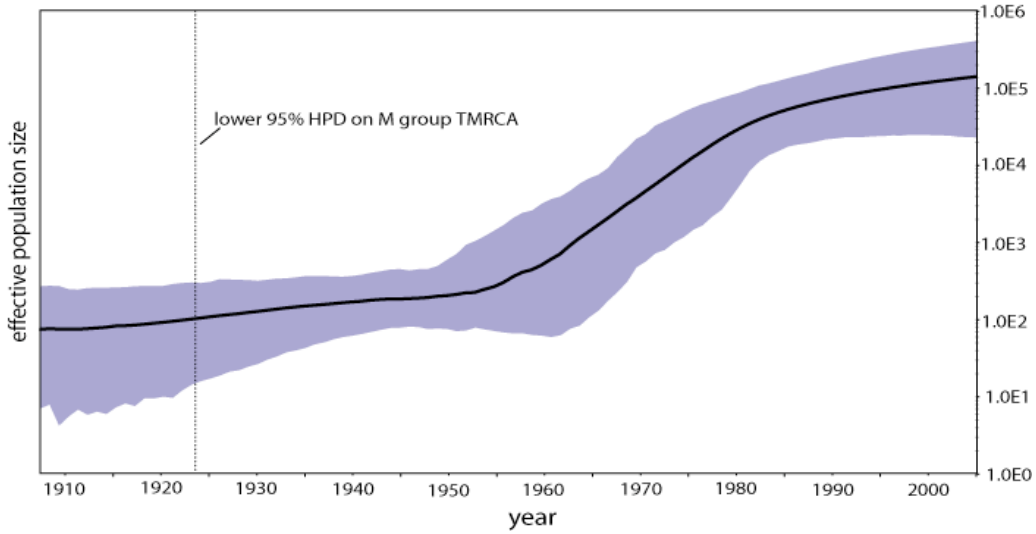
But...

- What population are we actually studying the dynamics of?
  - Pathogens? Then we are completely ignoring the massive population structure imposed by transmission
  - Infections? They don't reproduce in the way population genetics expects
- These are not models of disease transmission. The parameters (e.g. "effective population size") are hard to interpret in epidemiological terms to get incidence, prevalence, R_0, etc.
- Often a skyline plot is simply examined by eye for temporal and spatial trends without interpreting the actual numbers
- Nevertheless, these models have attractive simplicity and are easy to run

Example: emergence of HIV from Kinshasa

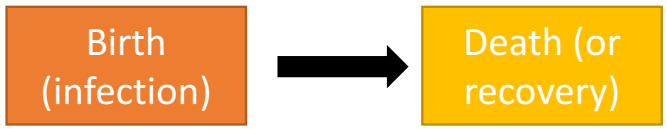
Worobey et al., *Nature*, 2008

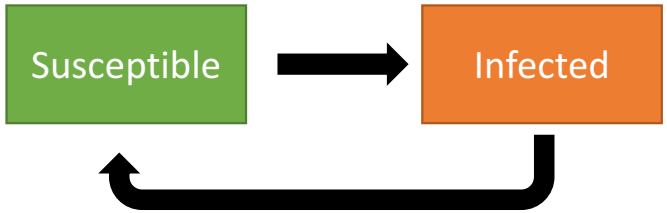




Forwards-time epidemiological models

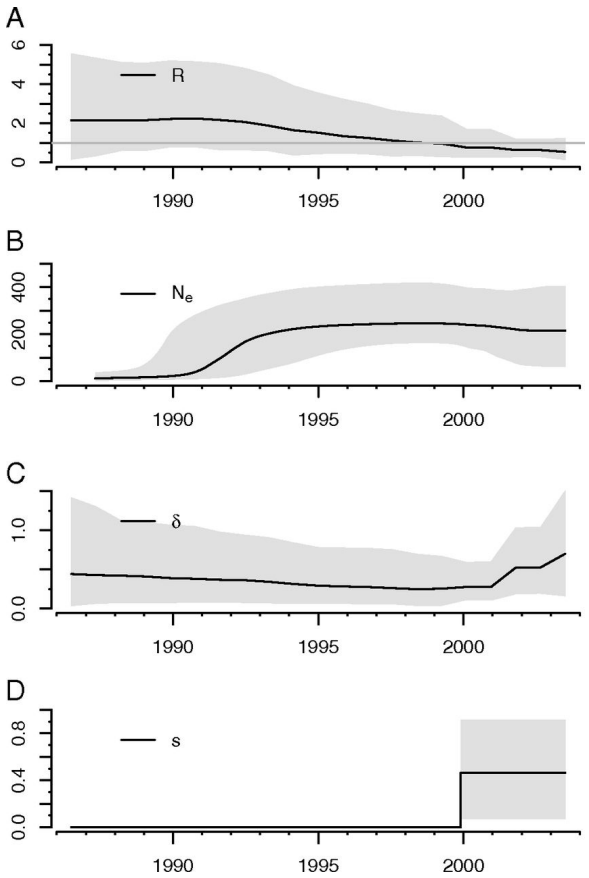
- Both coalescent models operate backwards in time
- Another family is forwards-time and behaves more like a conventional epidemiological model
- The sampling process must be modelled here along with transmission
  - Coalescent models deal only with the history of sample, so can avoid this
- Well-known models implemented in a phylodynamics framework:
  - Birth-death (Stadler et al., *Mol Biol Evol*, 2011)
  - Birth-death with time-varying parameters (Stadler et al., *PNAS*, 2013)
  - Birth-death with population structure (Kühnert et al., *Mol Biol Evol*, 2016)
  - SIS (Leventhal et al., *Mol Biol Evol*, 2014)
  - SIR (Kühnert et al., *J R Soc Interface*, 2014)



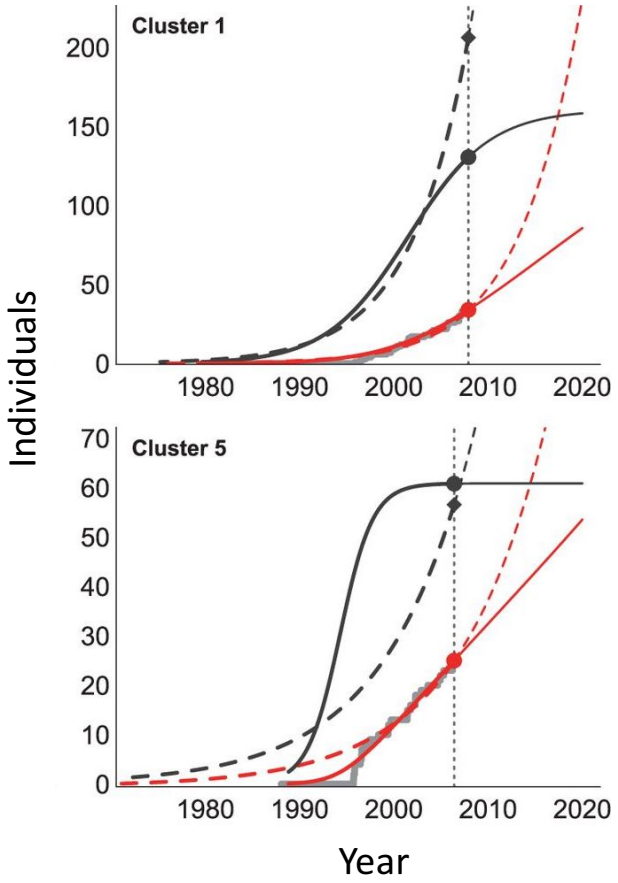


Examples: HIV in Europe

Stadler et al., *PNAS*, 2013 (UK data)



Leventhal et al., *Mol Biol Evol*, 2014 (Swiss data)

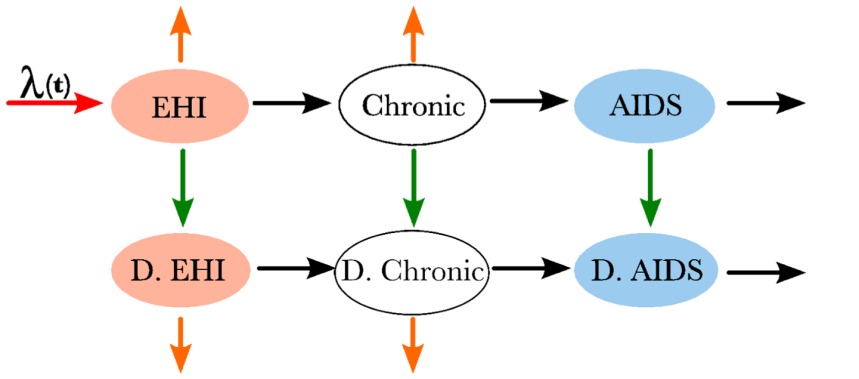


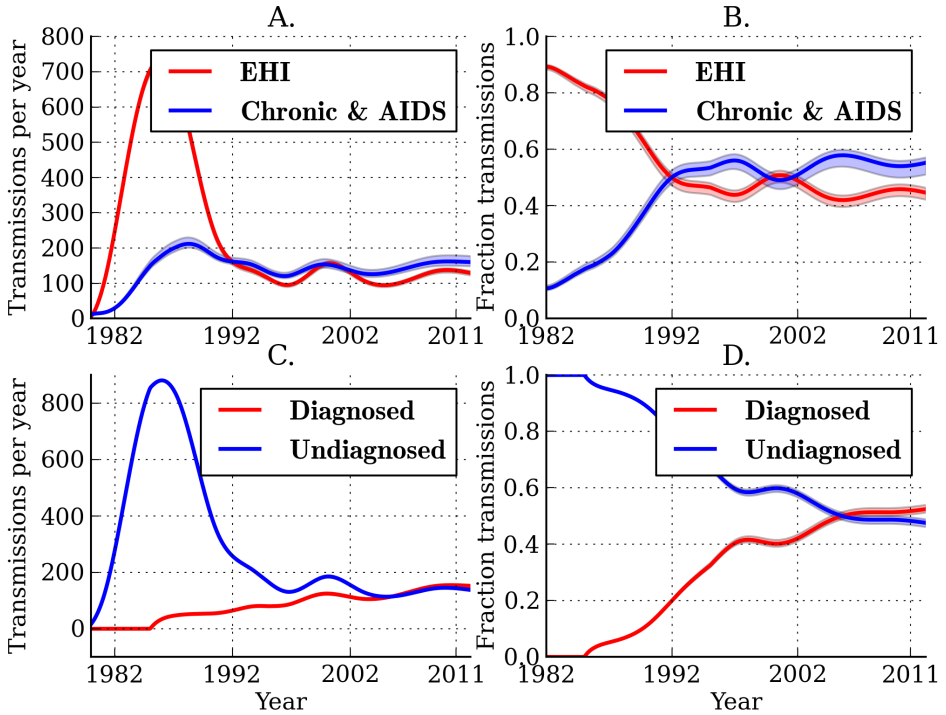
The epidemiological structured coalescent

- Coalescent models reformulated to mimic epidemic models
- The “compartments” of a classic epidemic model become the “demes” in the structured coalescent
  - Not necessarily geographical
- As these are backwards-time models, no need to model sampling
- Packages
  - phydyn (Volz and Siveroni, *PLOS Comput Biol*, 2018)
  - MASCOT (Müller et al., *Bioinformatics*, 2018)

Example: HIV transmission in early infection, Detroit MSM

Volz et al., *PLOS Med*, 2013



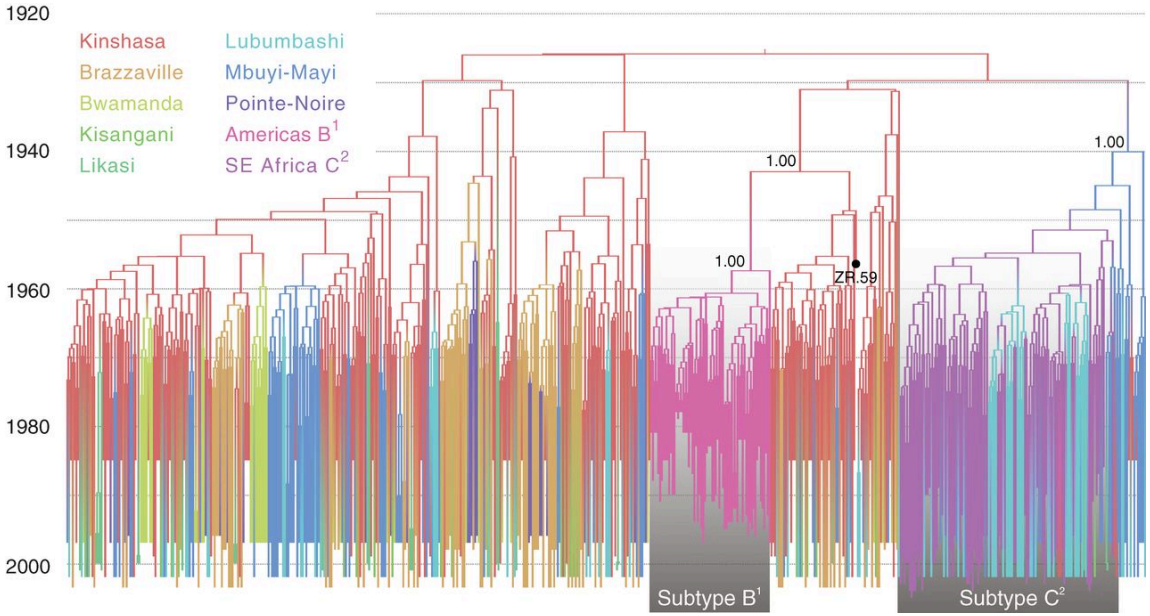


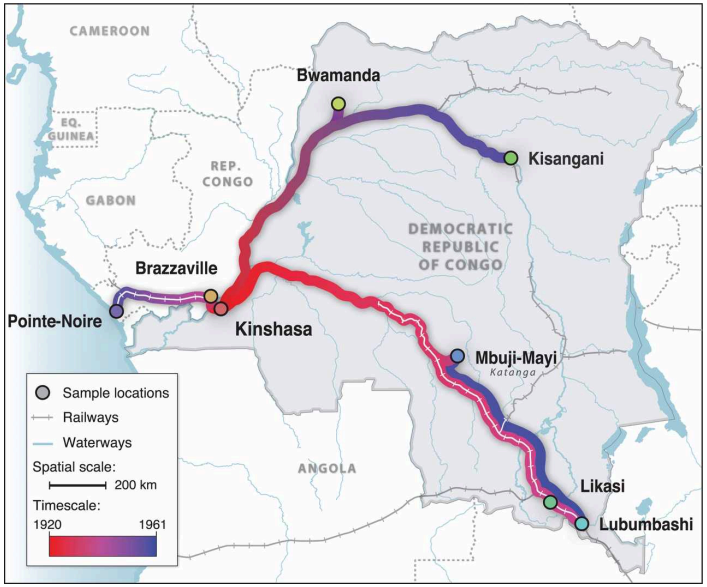
Phylogeography

- Phylogeography infers the movement of ancestral lineages through space and time using the phylogeny
  - Related is *phyloanatomy*, inferring movement between compartments of a host organism
- The structured coalescent is one approach to phylogeography, but it struggles with large numbers of demes, and also assumes a finite number of discrete locations
- The "mugration" model (e.g. Lemey et al., *PLOS Comput Biol*, 2009) treats location like a nucleotide and can handle many, many states
  - Caution: sampling bias
- A continuous model also exists (Lemey et al., *Mol Biol Evol*, 2010), if samples with exact latitude and longitude and available
- Key difference: the diffusion process is carried along the tree branches (like mutations) but is not assumed to generate it
  - Lineage splits are not part of the model

Example: HIV in the DRC

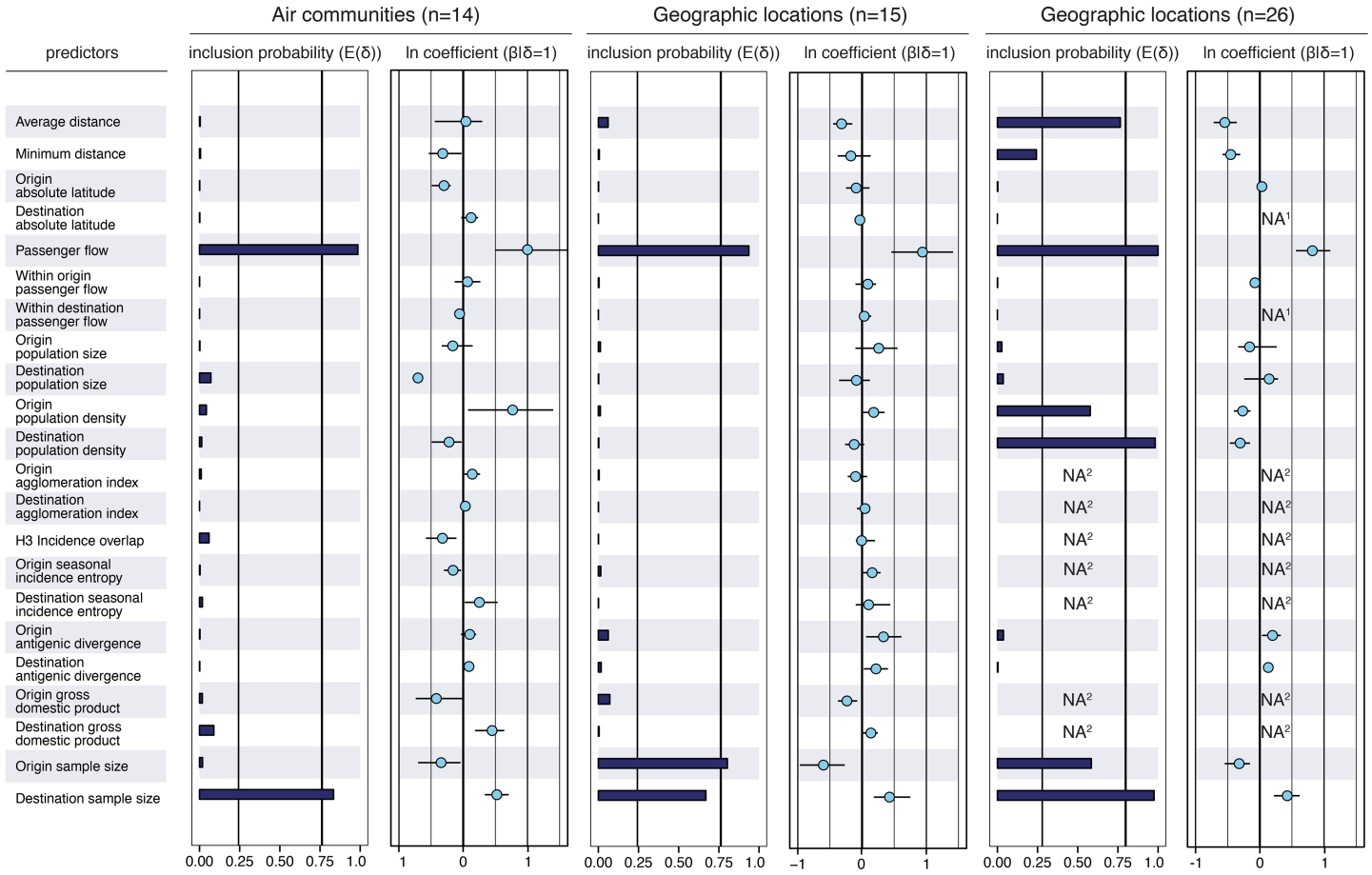
Faria et al., *Science*, 2014





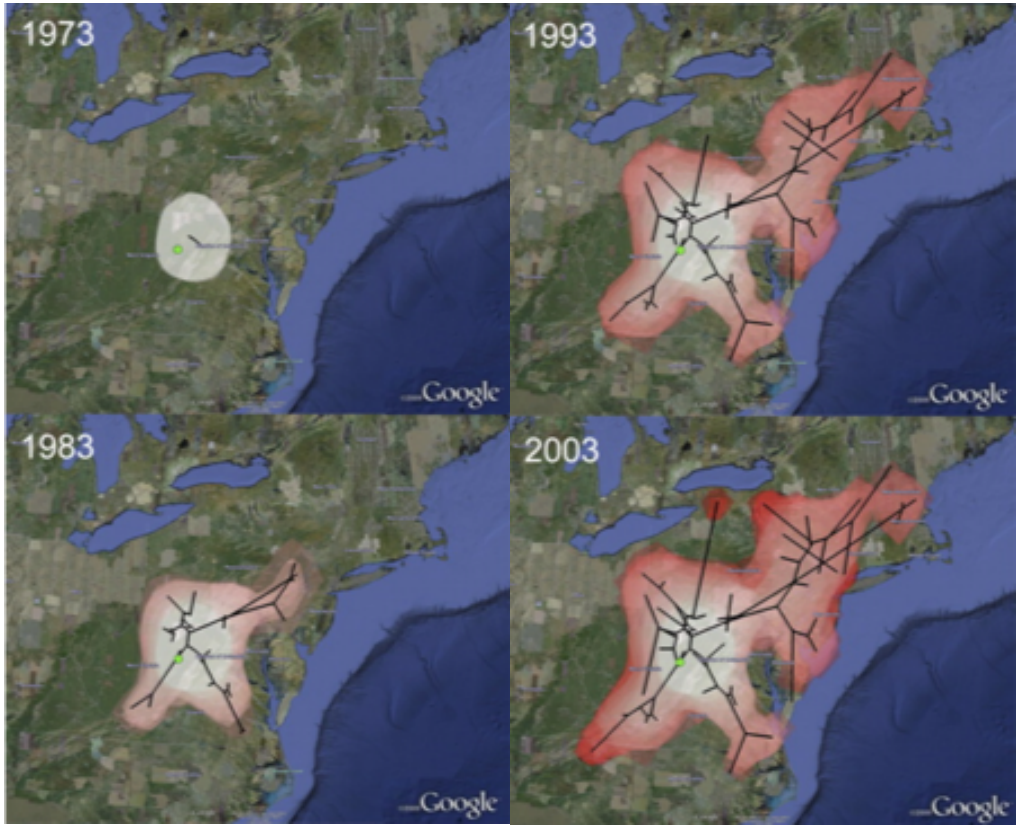
Example: predictors of influenza H3N2 spread

- Phylogeography models can also be used to determine significant predictors of transitions between locations
- Lemey et al., *PLOS Pathogens*, 2014.



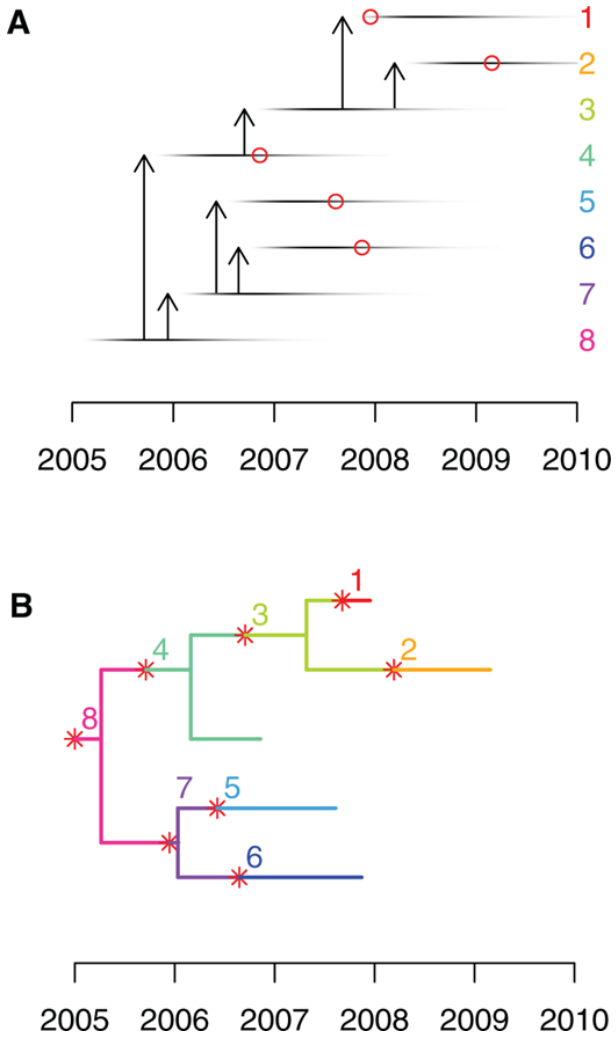
Example: raccoon rabies in North America

(Lemey et al., *Mol Biol Evol*, 2010)



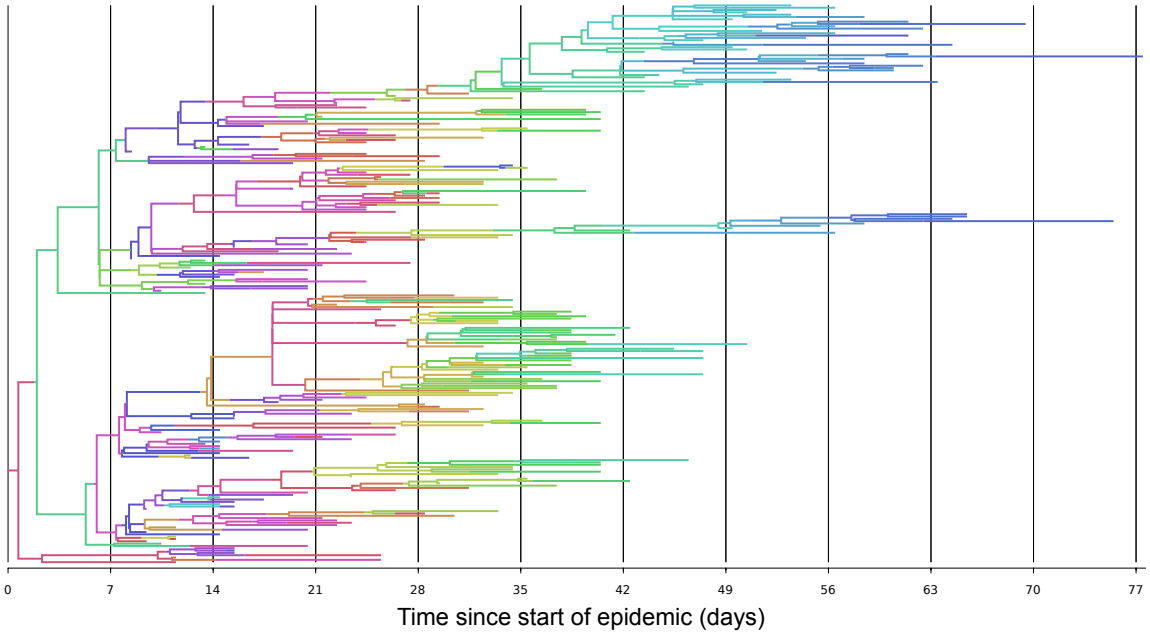
Transmission tree reconstruction

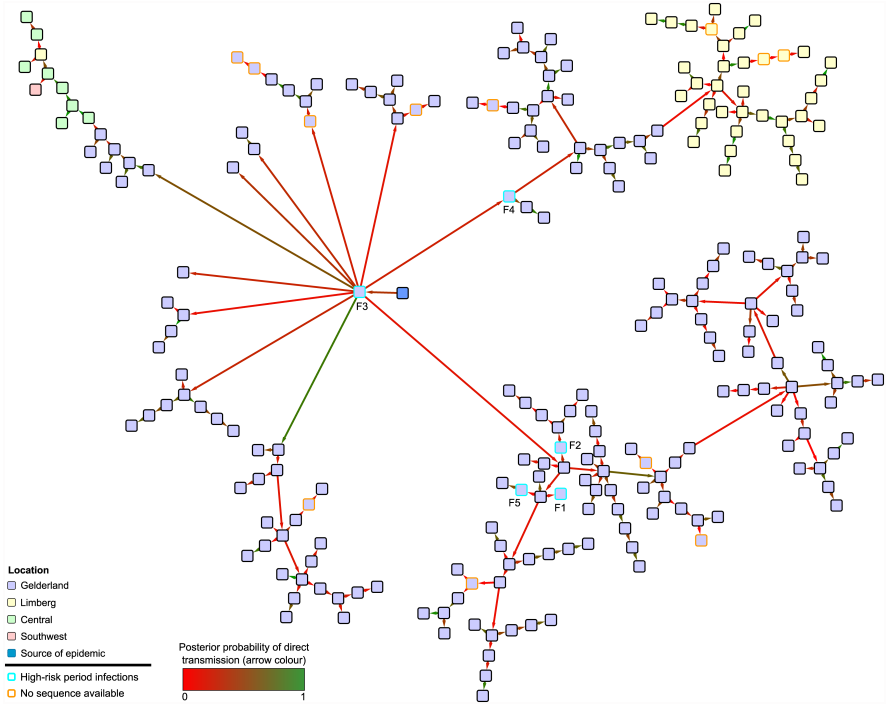
- The models discussed up until now are used to infer general properties of the epidemic (transmission rates, reproductive numbers, etc.)
  - Coalescent models assume sampling is sparse
  - Forwards-time models have an explicit sampling probability parameter
  - Neither are fundamentally concerned with exactly who infected who
- Transmission tree reconstruction (or source attribution) methods are concerned with exactly how the samples relate to each other in the transmission chain, rather than the properties of the entire epidemic
  - Many methods assume complete sampling
  - Not all are phylogenetic or phylodynamic
  - TransPhylo (Didelot et al., *Mol Biol Evol*, 2014 & 2017), BEASTLIER (Hall et al., *PLOS Comput Biol*, 2015), phybreak (Klinkenberg et al., *PLOS Comput Biol*, 2017), structured coalsecent source attribution (Volz and Frost, *PLOS Comput Biol*, 2013), SCOTTI (de Maio et al., *PLOS Comput Biol*, 2016)
  - Phyloscanner?



(Didelot et al., *Mol Biol Evol*, 2017)

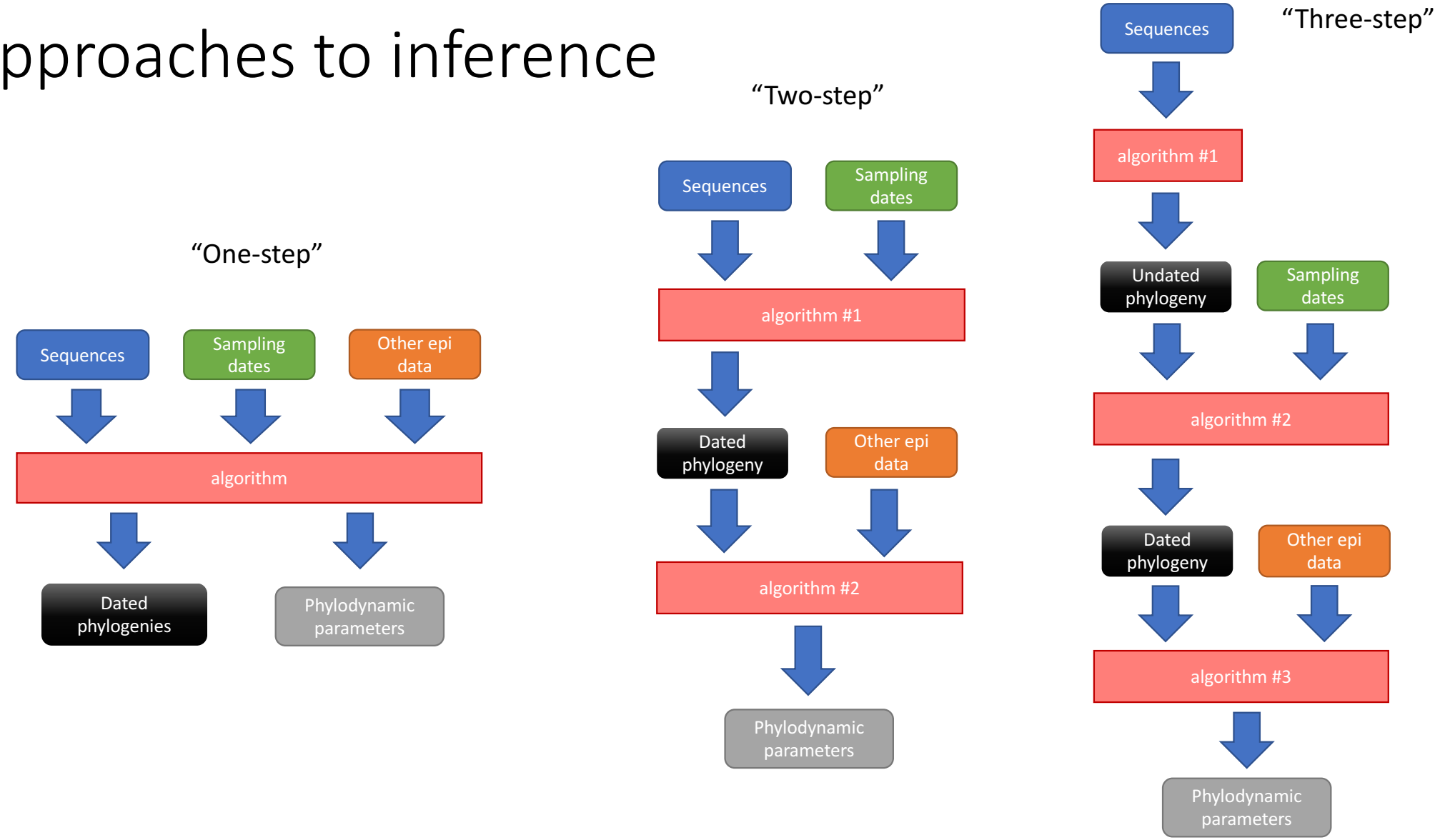
Example: H7N7 avian influenza, Netherlands, 2003





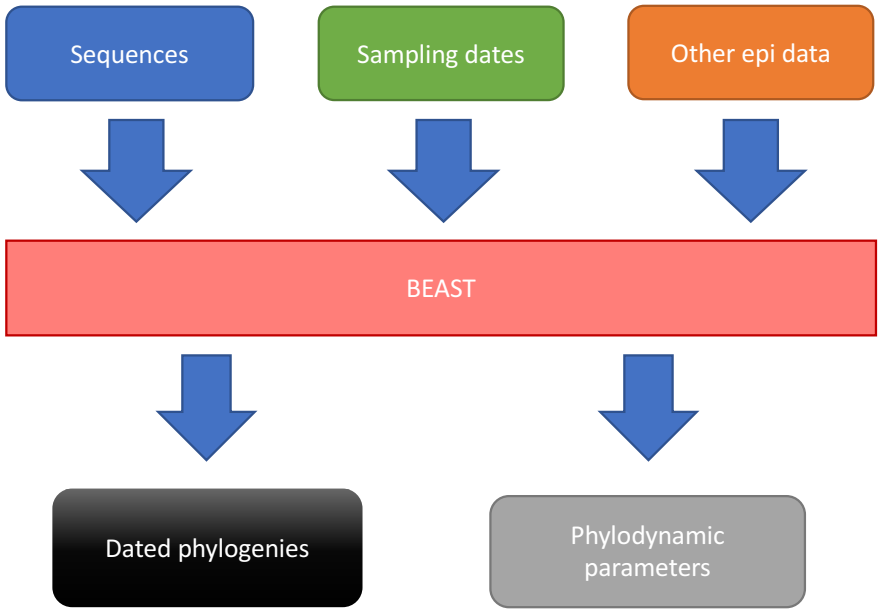
(Hall et al., *PLOS Comput Biol*, 2015)

Approaches to inference



BEAST

- **B**ayesian **E**volutionary **A**nalysis (by) **S**ampling **T**rees
- The current gold standard one-step package(s)
  - Deals naturally with issues surrounding phylogenetic uncertainty
  - ...but struggles with datasets beyond a few hundred sequences, especially for more complex models
- BEAST analyses are usually presented as taking sequences as input (one-step), but it can also use a fixed tree (two- or three-step)



The octopus and the mouse

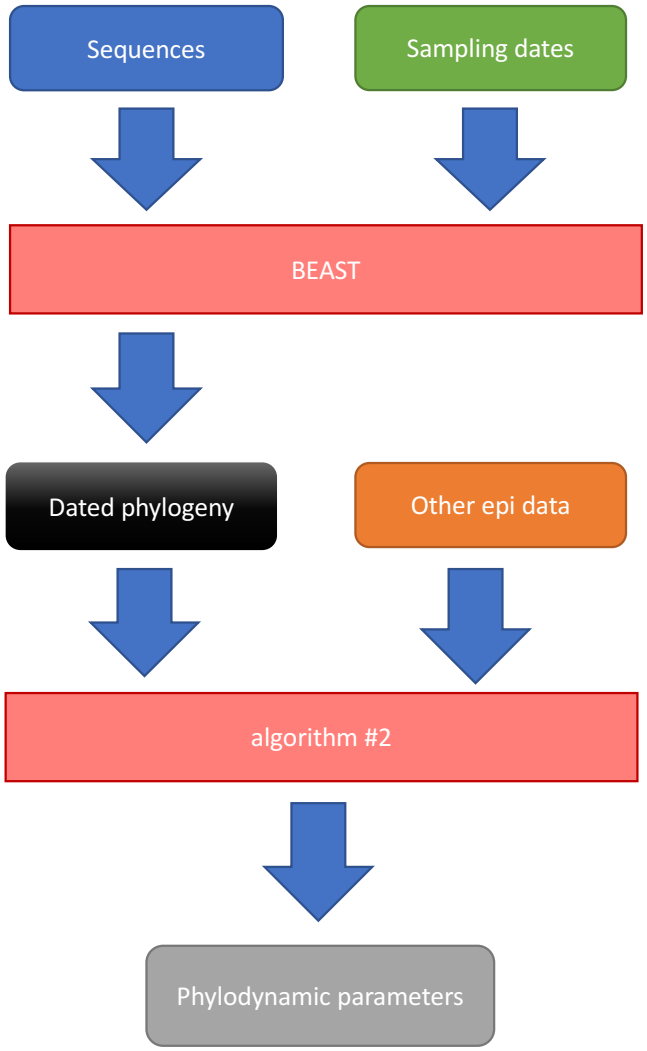
- Confusingly, there are two BEASTs which are independent development projects. They can be used interchangeably for many, but not all analyses

- Suchard et al., *Virus Evol*, 2018 (most recent citation)
- Arguably more user-friendly
- Cutting-edge for population-genetic coalescent models and phylogeography

- Bouckaert et al., *PLOS Comput Biol*, 2014
- More flexible, modular structure
- Cutting-edge for epidemiological models (coalescent and forwards-time)

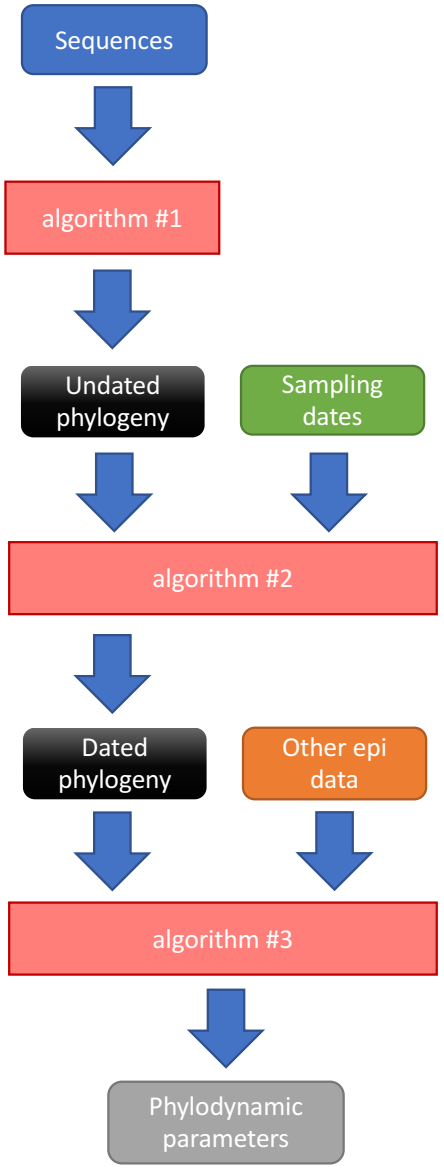
Two-step analysis

- If a standard BEAST analysis will run, but your phylodynamic model will not (or is not in BEAST):
  1. Build the dated phylogeny or phylogenies with BEAST
  2. Run a separate algorithm for phylodynamic inference
     - Sometimes this too is BEAST!



Beyond the BEAST limits

- If your dataset is so large that BEAST will not converge in reasonable time, you need the three-step process
  1. Make a phylogeny with branch lengths in genetic units using a standard package (usually maximum likelihood)
  2. Use a separate package to infer a molecular clock and fit the tree to a timeline
  3. Use that dated phylogeny to fit the phylodynamic model
- Now several options for step 2:
  - Least-squares dating (LSD) (C++; To et al., *Syst Biol*, 2016)
  - node.dating (R; Jones and Poon, *Bioinformatics*, 2017)
  - TreeTime (Python; Sagulenko et al., *Virus Evol*, 2017)
  - treedater (R; Volz and Frost, *Virus Evol*, 2017)
  - BactDating (R; Didelot et al., *Nucleic Acids Res*, 2018)
- Limited scope for dealing with phylogenetic uncertainty



Summary

- Phylodynamics marries evolutionary biology and mathematical modelling of infectious disease
- Phylogenies with branch lengths in calendar time are almost always used
- The phylogeny is taken to be a history of an epidemic, and by fitting models to that history, we recover important parameters of that epidemic
- Phylogeography and source attribution are related areas
- Bayesian methods are most common
  - One-step BEAST for datasets up to a few hundred sequences and established models
  - Two-step procedures for more experimental models
  - Three-step procedures for large datasets