Phylodynamics: phylogenetics meets epidemiological modeling

Molecular Epidemiology of Infectious Diseases

Lecture 12

April 6th, 2026

The road here

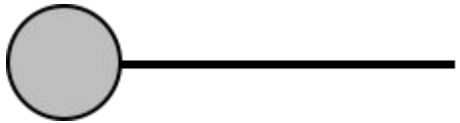
Two weeks ago: Modeling epidemic dynamics with SIR models

Last week: Stochastic models for simulation and inference

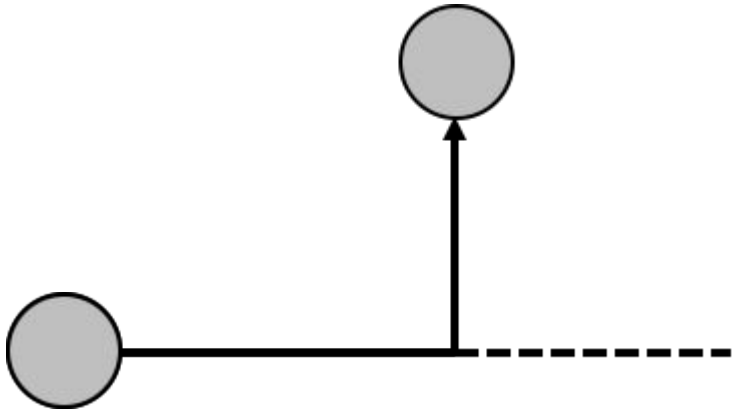
This week: putting everything together with phylodynamic modeling

Now we get to put all the pieces back together again!

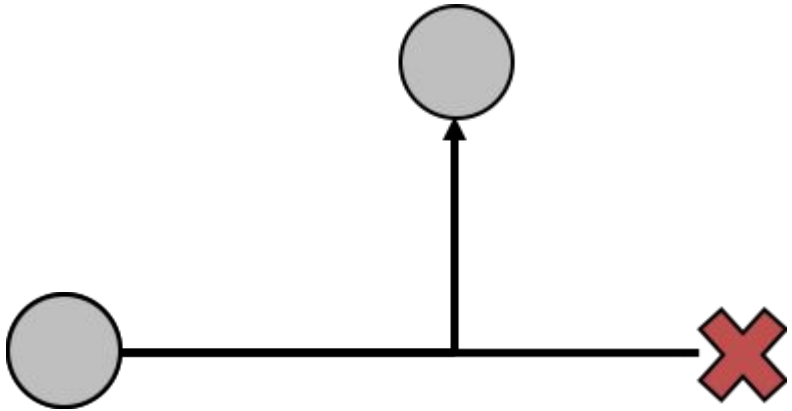
A simple epidemic example



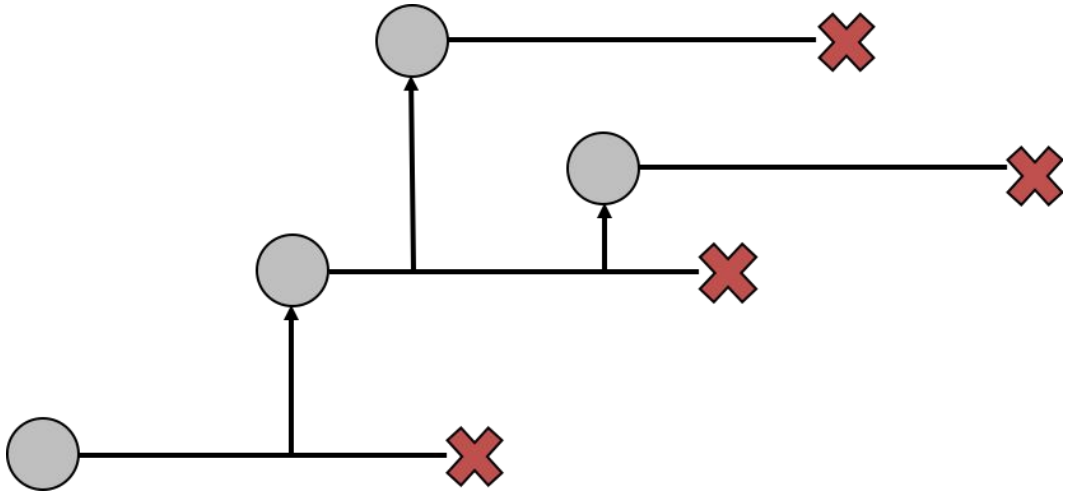
A simple epidemic example



A simple epidemic example

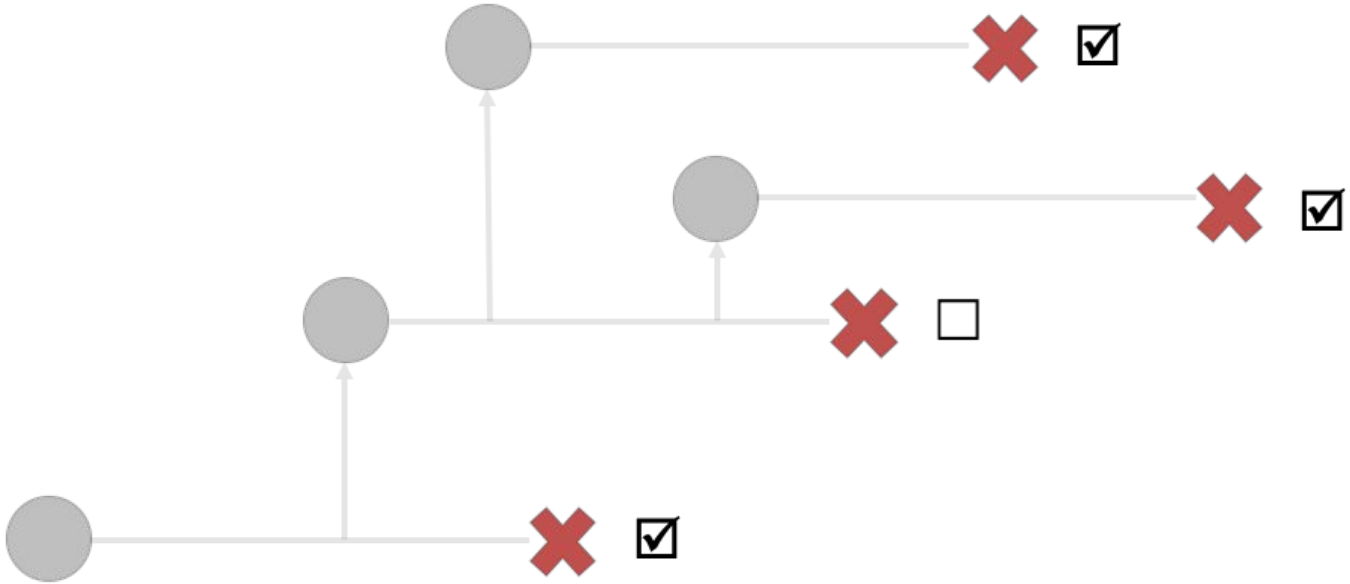


A simple epidemic example



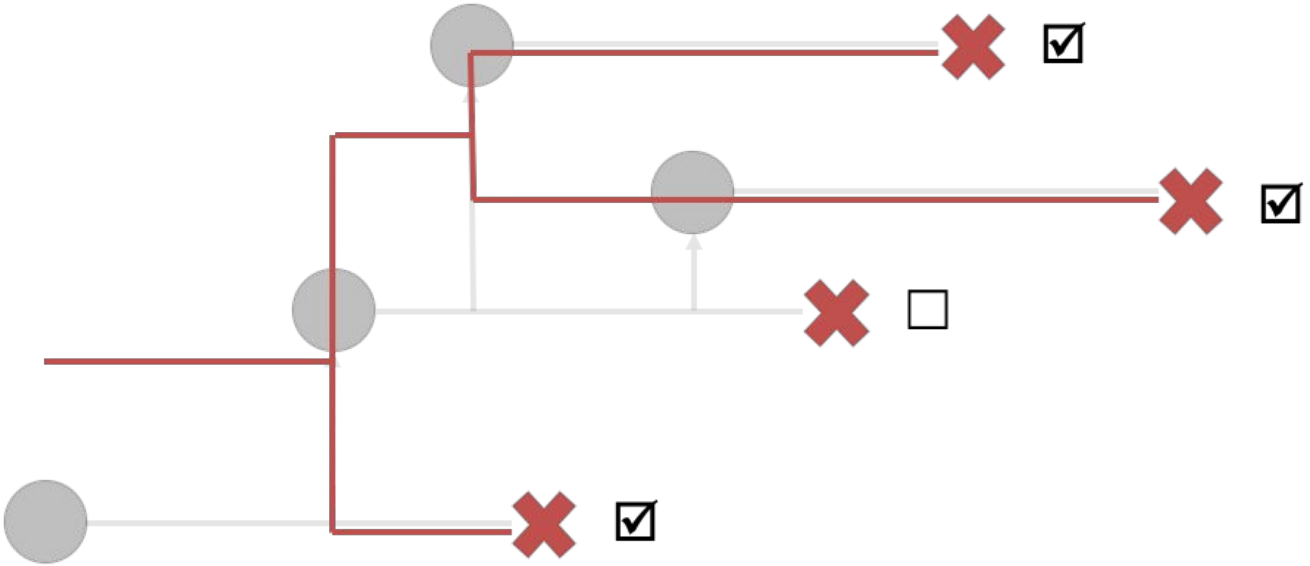
Transmission tree

A simple epidemic example

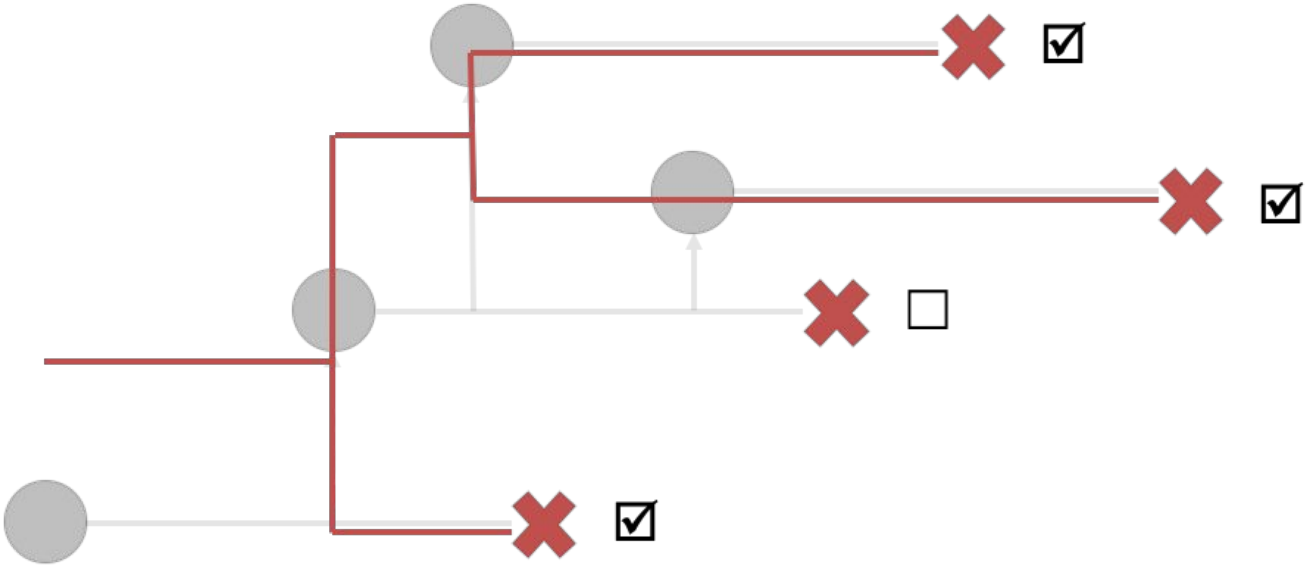


Transmission tree with incomplete sampling

A simple epidemic example



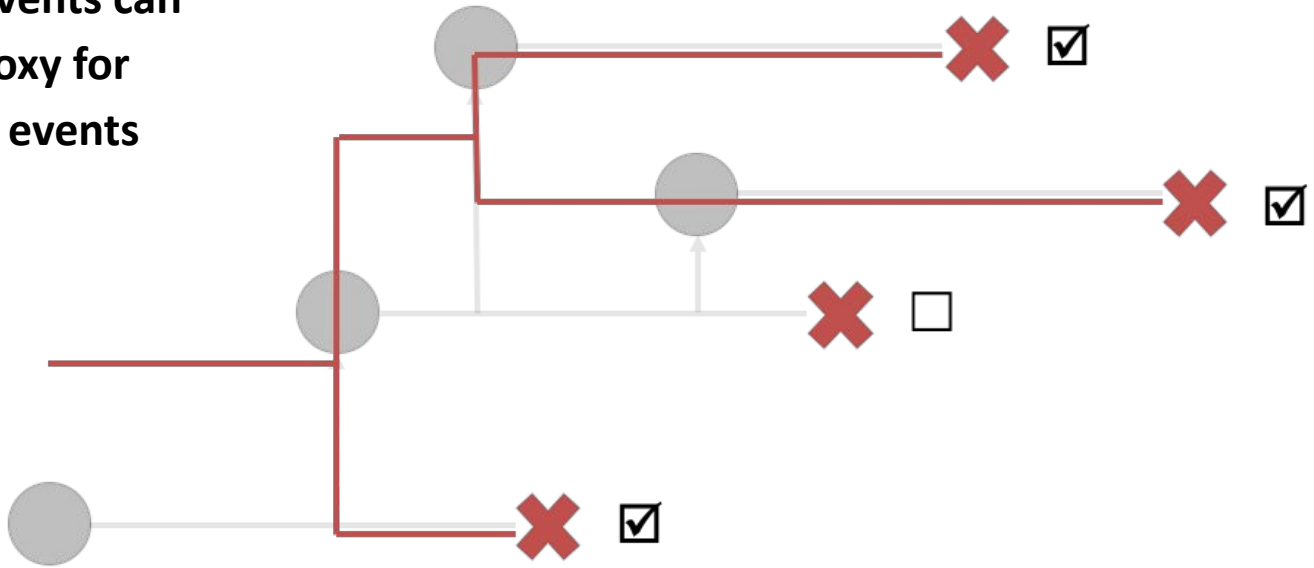
A simple epidemic example



Pathogen phylogenies recapitulate major features of the underlying transmission tree

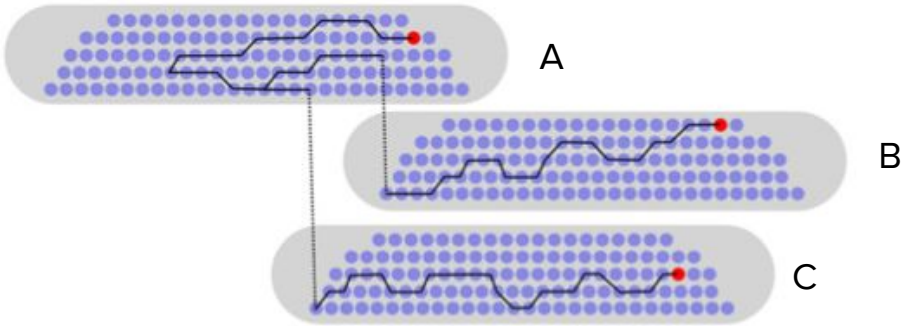
A simple epidemic example

Coalescent events can serve as a proxy for transmission events



Within-host diversity

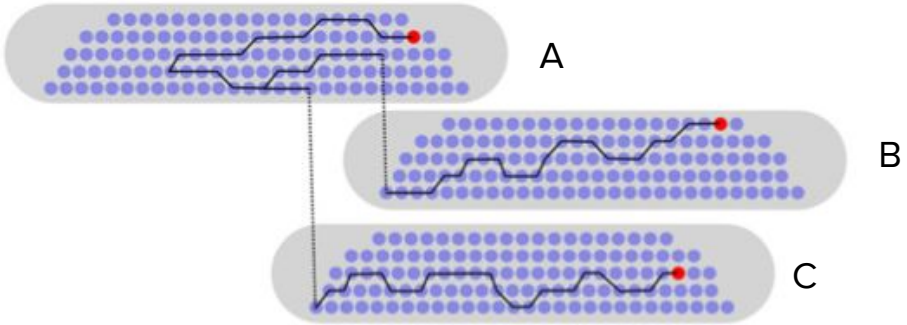
The exact branching structure of the phylogeny will depend on the timing and order of coalescent events within hosts...

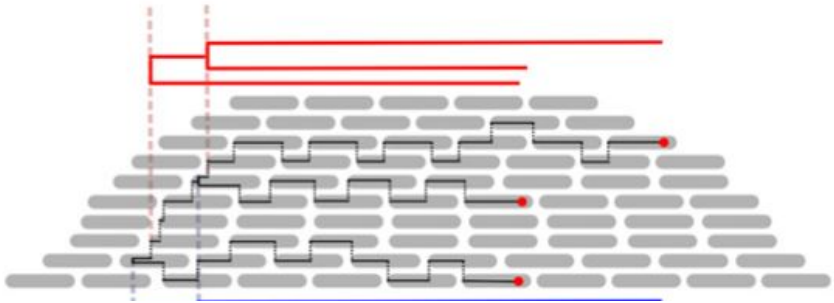


Ympa *et al.* (Genetics, 2013)

The macroscopic view

If we assume a single pathogen lineage resides in each infected hosts, coalescent events will closely coincide with transmission events.



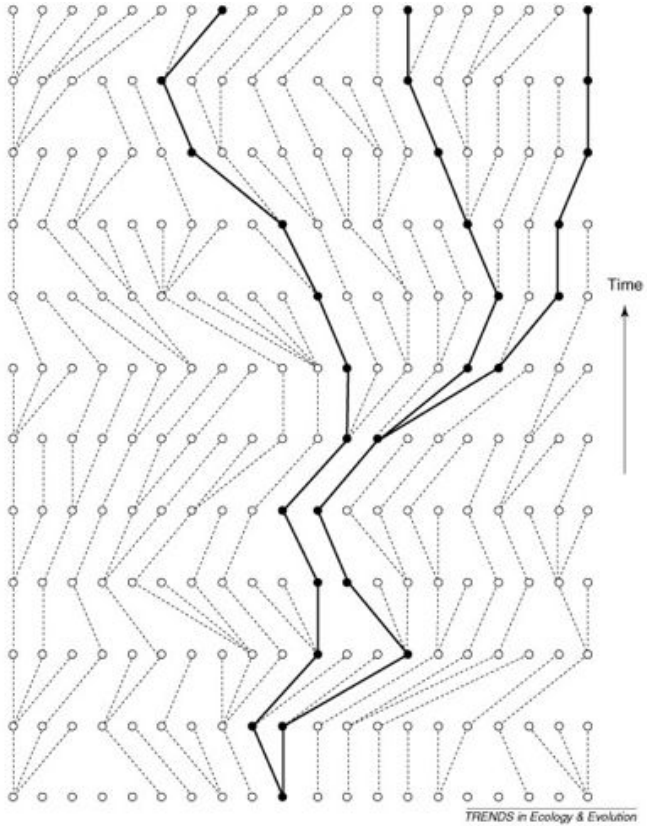


Ympa *et al.* (Genetics, 2013)

Basic coalescent theory

The probability of two lineages coalescing per generation is:

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



Kuhner *et al.* (2008)

Basic coalescent theory

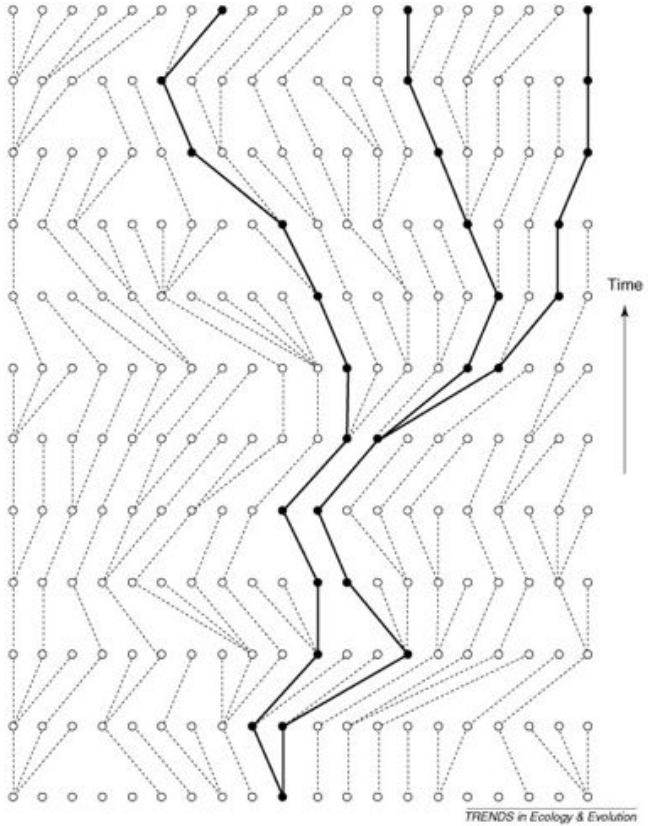
The probability of two lineages coalescing per generation is:

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

The rate of coalescence in continuous time:

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

$$Pr(X = t) = \lambda e^{-\lambda t}$$

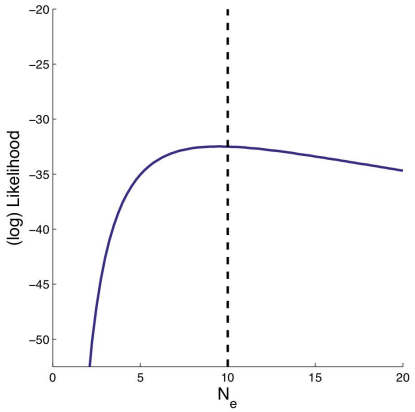


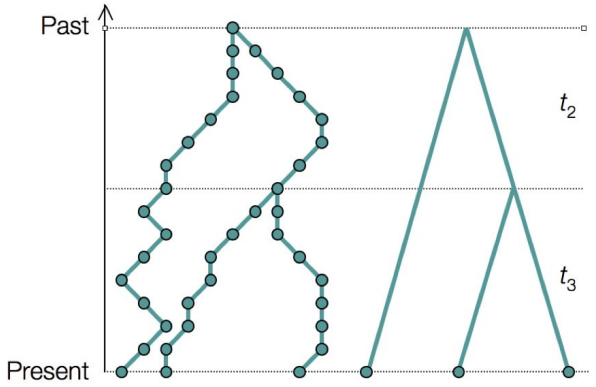
Kuhner *et al.* (2008)

Coalescent-based inference

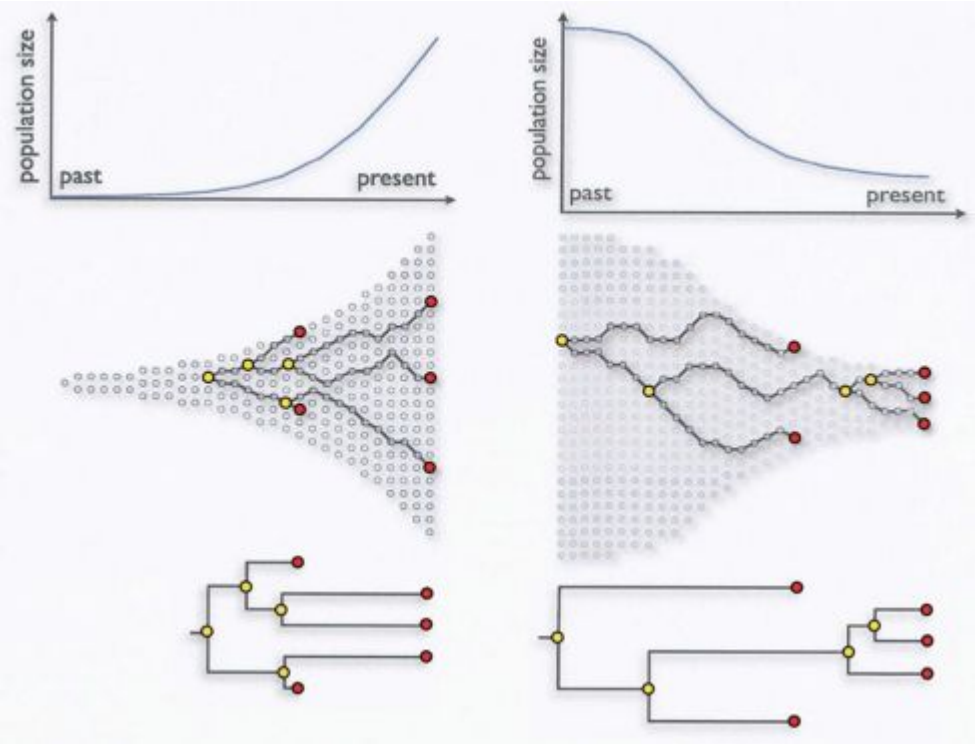
We can therefore infer demographic parameters like N_e from a known phylogeny.

$$L(T|N_e) = \frac{1}{N_e^{(n-1)}} \prod_{k=2}^{n} \exp\left(- \frac{\binom{k}{2}}{N_e} t_k \right)$$





The signal of population size change

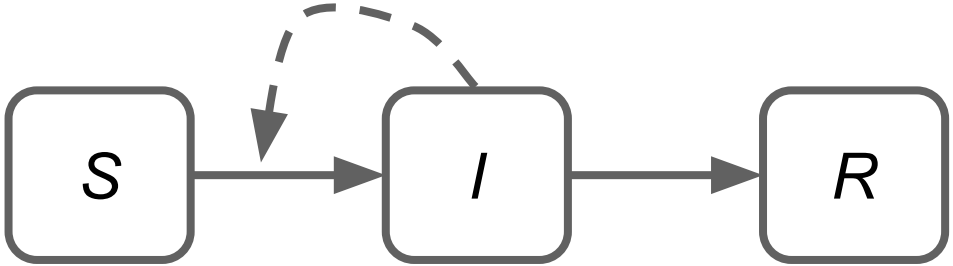


But are these appropriate coalescent models for an infectious pathogen?

The SIR model

Under the SIR model the transmission or “birth rate” of new pathogen lineages depends on the incidence which depends on both S and I.

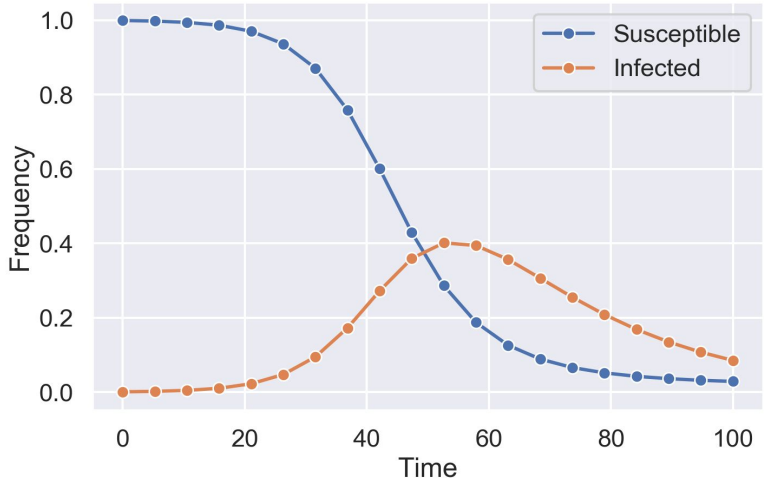
$$\frac{dS}{dt} = -\beta SI$$
$$\frac{dI}{dt} = \beta SI - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$



The SIR model

Under the SIR model the transmission or “birth rate” of new pathogen lineages depends on the incidence which depends on both S and I.

$$\begin{aligned}
\frac{dS}{dt} &= -\beta SI \\
\frac{dI}{dt} &= \beta SI - \gamma I \\
\frac{dR}{dt} &= \gamma I
\end{aligned}$$



The SIR coalescent model

Starting from first principles, we can intuitively reason that ***two things must occur*** in order for a pair of lineages (currently residing in two different infected hosts) to coalesce:

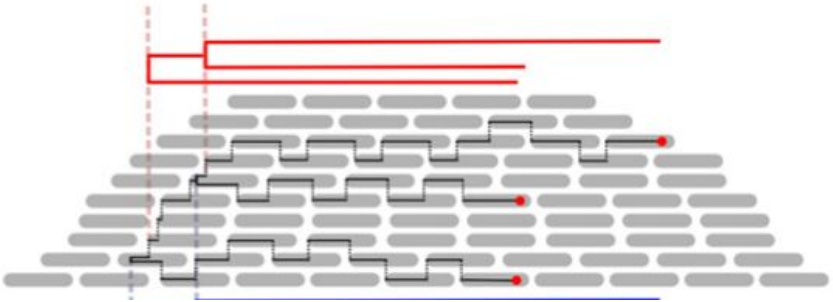
1. A transmission event must occur somewhere in the host population.

2. Our pair of lineages must reside in the two infected hosts resulting from the transmission event.

The SIR coalescent model

Starting from first principles, we can intuitively reason that ***two things must occur*** in order for a pair of lineages (currently residing in two different infected hosts) to coalesce:

1. A transmission event must occur somewhere in the host population.



SIR-type coalescent model

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

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

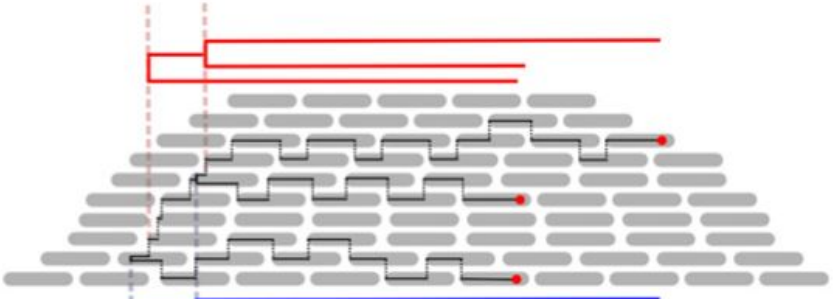
Volz *et al.* (Genetics, 2009)
Volz (Genetics, 2012)

The SIR coalescent model

Starting from first principles, we can intuitively reason that ***two things must occur*** in order for a pair of lineages (currently residing in two different infected hosts) to coalesce:

2. Our pair of lineages must reside in the two infected hosts resulting from the transmission event.

$$S + I \xrightarrow[\beta]{} 2I$$



Probability of two lineages coalescing

What is the probability that one pair of lineages coalesces rather than any other pair of lineages?

There are *I(t)* total pathogen lineages in the host population at time *t*.

The total number of pairs that could possibly coalesce is:

$$\text{Total pairs} = \binom{I(t)}{2} = \frac{I(t)\big(I(t)-1\big)}{2}$$

The probability that our pair of lineages coalesces is therefore:

$$p_{coal} = \frac{1}{\text{Total pairs}} = \frac{2}{I(t)\big(I(t)-1\big)} \approx \frac{2}{I(t)^2}$$

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:

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

The pairwise coalescent rate is therefore:

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

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

Comments on the SIR coalescent model

As before, the rate of coalescence is still inversely proportional to prevalence (infected population size).

But now the rate at which two pathogen lineages coalesce not only depends on prevalence but also on incidence (transmission rates)!

This means that the coalescent rate will peak when incidence is high but prevalence is low.

What makes these models powerful is that we can now fit epidemiological models directly to pathogen phylogenies.

Host population structure

Host populations are almost always structured into different subpopulations:

- Age structure
- Risk/contact structure
- Spatial structure
- Stages of disease progression
- Multiple host/vector species

The problem with population structure

Standard coalescent models assume that all lineages in the tree are **exchangeable**.

Exchangeability here means that any lineage is equally likely to coalesce with any other lineage in the tree.

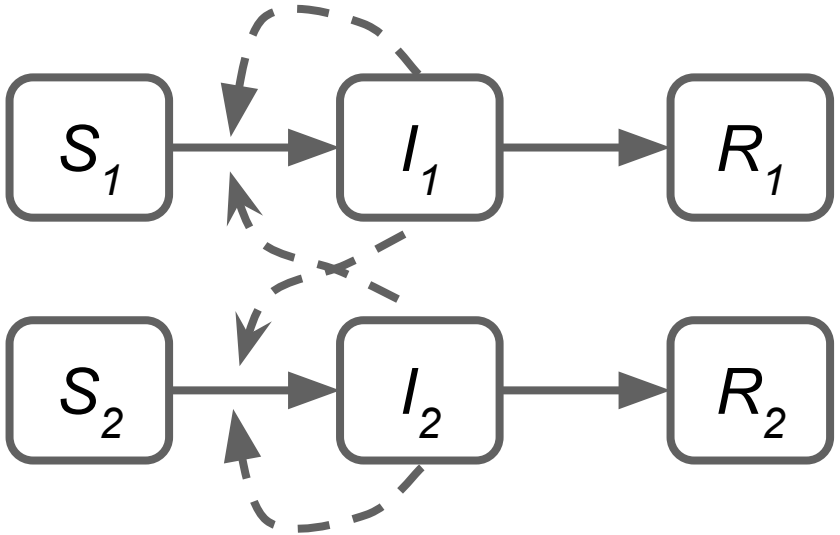
Many forms of population structure violate this key assumption.

We therefore need structured epidemiological/coalescent models!

SIR model with multiple host classes

The SIR model can be generalized to include multiple different “compartments”:

$$\frac{dI_i}{dt} = \sum_j \beta_{ji} S_i I_j - \gamma I_i$$



The Volz Structured Coalescent

The structured coalescent model of Volz (Genetics, 2012) considers both complex epidemiological dynamics and host population structure!

Under this model, the pairwise coalescent rate for two lineages *i* and *j*:

$$\lambda_{ij} = \sum_{k}^{m} \sum_{l}^{m} \frac{f_{kl}}{y_k y_l} \left(p_{ik} p_{jl} + p_{il} p_{jk} \right)$$

The Volz Structured Coalescent

$$\lambda_{ij} = \sum_{k}^{m} \sum_{l}^{m} \frac{f_{kl}}{y_k y_l} \left(p_{ik} p_{jl} + p_{il} p_{jk} \right)$$

Sum over all possible locations of both lineages

The Volz Structured Coalescent

Birth or transmission events

$$\lambda_{ij} = \sum_{k}^{m} \sum_{l}^{m} \frac{f_{kl}}{y_k y_l} \left(p_{ik} p_{jl} + p_{il} p_{jk} \right)$$

The Volz Structured Coalescent

$$\lambda_{ij} = \sum_{k}^{m} \sum_{l}^{m} \frac{f_{kl}}{y_k y_l} \left(p_{ik} p_{jl} + p_{il} p_{jk} \right)$$

Pairs of individuals in populations *k* and *l*

The Volz Structured Coalescent

p_{ik} is the probability that lineage *i* is in state *k*

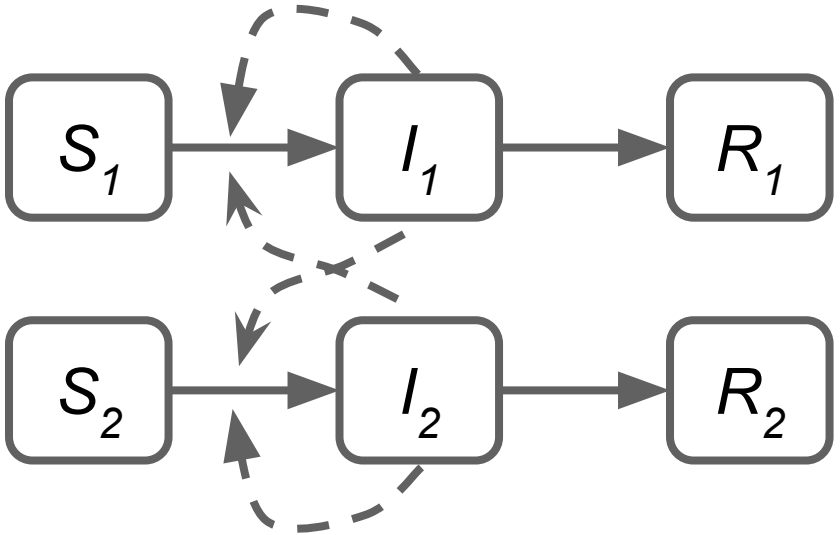
$$\lambda_{ij} = \sum_{k}^{m} \sum_{l}^{m} \frac{f_{kl}}{y_k y_l} \left(p_{ik} p_{jl} + p_{il} p_{jk} \right)$$

Lineage state probabilities

SIR coalescent model with structure

The Volz structured coalescent model for the SIR model with multiple host classes:

$$\lambda_{ij} = \sum_{k}^{m} \sum_{l}^{m} \frac{\beta_{kl} S_l I_k}{I_k I_l} \left(p_{ik} p_{jl} + p_{il} p_{jk} \right)$$



The Volz Structured Coalescent

The structured coalescent model of Volz (Genetics, 2012) considers both complex epidemiological dynamics and host population structure!

Under this model, the pairwise coalescent rate for two lineages *i* and *j*:

$$\lambda_{ij} = \sum_{k}^{m} \sum_{l}^{m} \frac{f_{kl}}{y_k y_l} \left(p_{ik} p_{jl} + p_{il} p_{jk} \right)$$

The location of of each lineage back is ***tracked probabilistically*** back through time in terms of the lineage state probabilities p_{ik}

The Volz Structured Coalescent

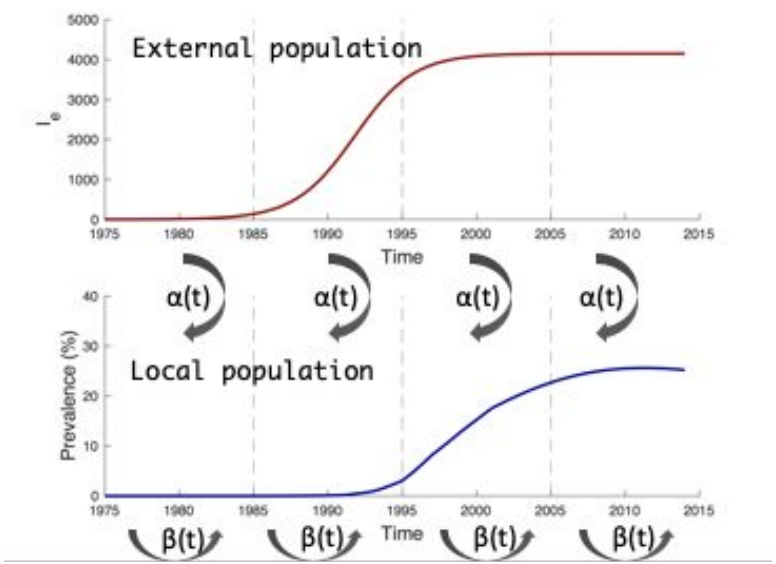
Lineage state probabilities p_{ik} are tracked backwards in time using a system of master equations (ODEs) based on the transition rates g_{kl}:

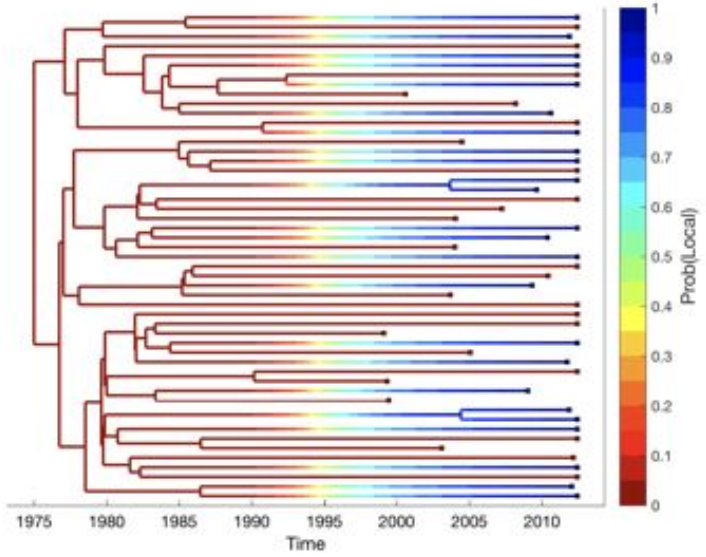
$$\frac{d}{dt}p_{ik} = \sum_{l}^{m} \left(p_{il} g_{kl} - p_{ik} g_{lk} \right)$$

Solving for p_{ik} at any point in time gives the probability that lineage *i* is in state *k*.

Volz (Genetics, 2012)

Tracking lineage probabilities for a two-location SIR model





Deriving other SIR-type models

The appropriate coalescent model for nearly any type of SIR model can be easily be derived from the Volz structured coalescent.

In the general case, there are two different ways lineages can move between populations:

- Through transmission events which can result in a coalescent event
- Through migration events involving a single lineage (no coalescence).

Deriving other SIR-type models

We will use the matrix ***F(t)*** to represent the rate at which lineages can coalesce and move between populations due to birth or transmission events:

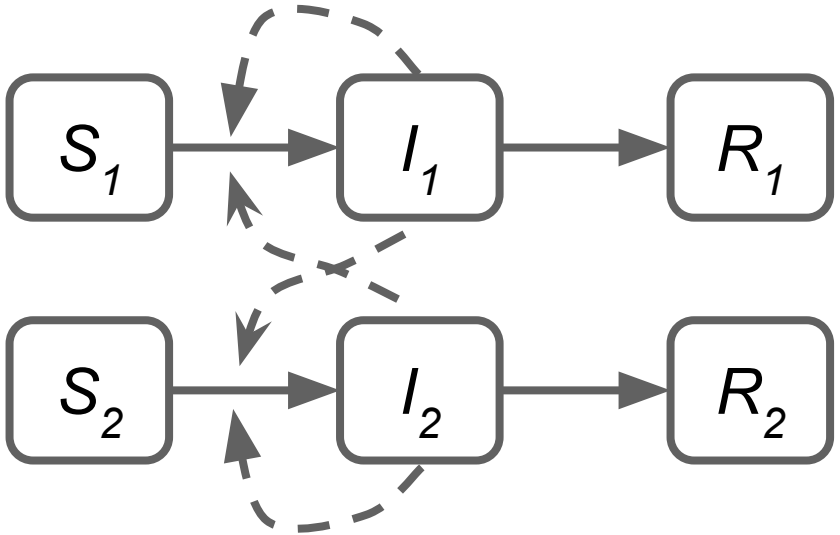
$$\boldsymbol{F}(t) = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{nn} \end{bmatrix}$$

We will use the matrix ***G(t)*** to represent the rate at which single lineages migrate between populations:

$$\boldsymbol{G}(t) = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ g_{n1} & g_{n2} & \cdots & g_{nn} \end{bmatrix}$$

Example: SIR model with two classes

$$\frac{dI_1}{dt} = \beta_{11} S_1 I_1 + \beta_{21} S_1 I_2 - \gamma I_1$$
$$\frac{dI_2}{dt} = \beta_{22} S_2 I_2 + \beta_{12} S_2 I_1 - \gamma I_2$$

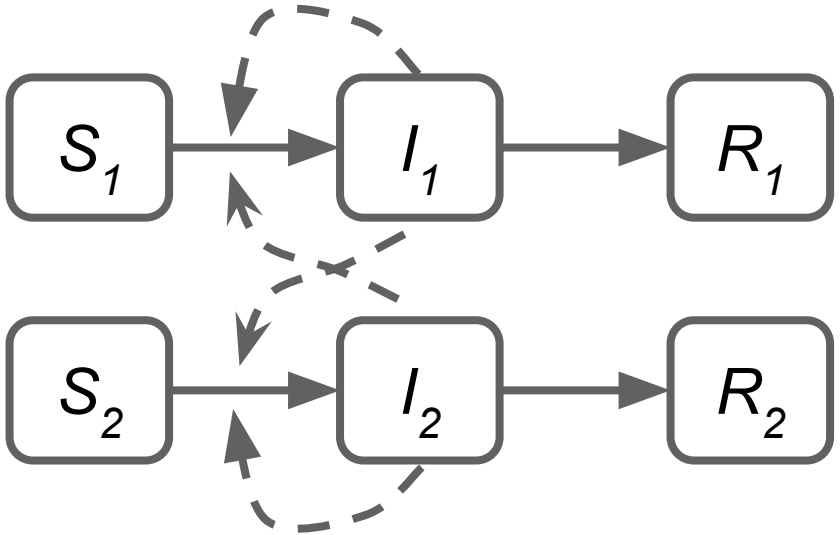


The SIR matrix decomposition

If we assume infected hosts do not migrate between populations:

$$\boldsymbol{F}(t) = \begin{bmatrix} \beta_{11} S_1 I_1 & \beta_{12} S_2 I_1 \\ \beta_{21} S_1 I_2 & \beta_{22} S_2 I_2 \end{bmatrix}$$

$$\boldsymbol{G}(t) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

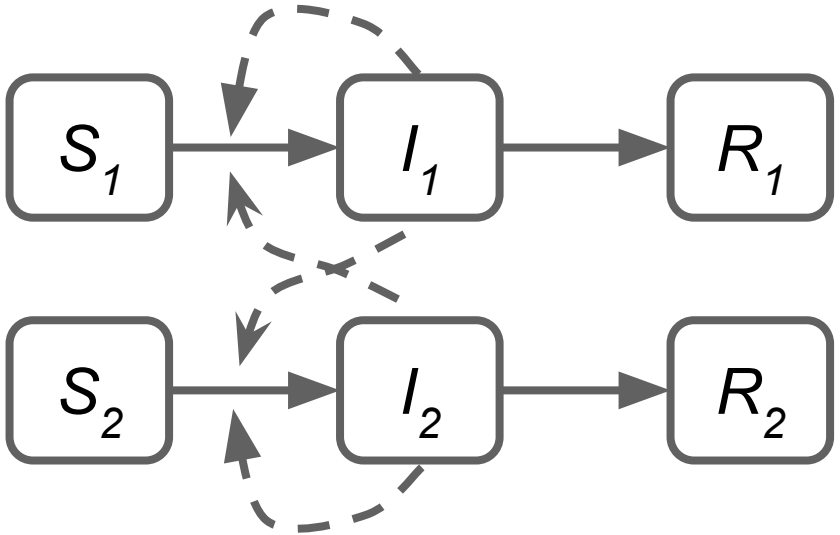


The SIR matrix decomposition

Or if we allow infected hosts to migrate between populations at rate γ_{ij}:

$$\boldsymbol{F}(t) = \begin{bmatrix} \beta_{11} S_1 I_1 & \beta_{12} S_2 I_1 \\ \beta_{21} S_1 I_2 & \beta_{22} S_2 I_2 \end{bmatrix}$$

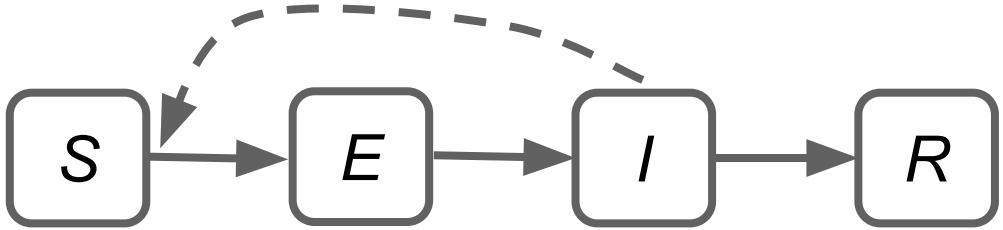
$$\boldsymbol{G}(t) = \begin{bmatrix} 0 & \gamma_{12} \\ \gamma_{21} & 0 \end{bmatrix}$$



Example: SEIR model

The SEIR model adds an exposed compartment to the basic SIR model:

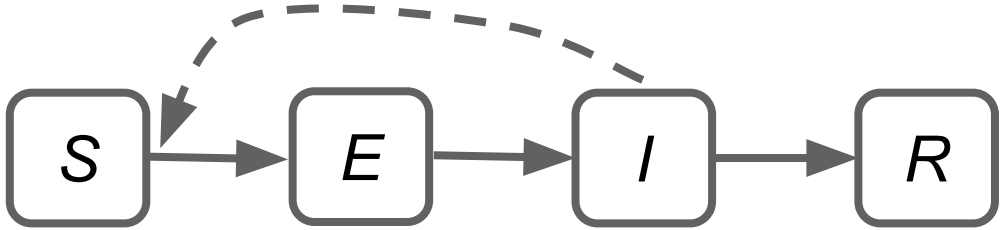
$$
\begin{aligned}
\frac{dS}{dt} &= -\beta SI \\
\frac{dE}{dt} &= \beta SI - \eta E \\
\frac{dI}{dt} &= \eta E - \gamma I \\
\frac{dR}{dt} &= \gamma I
\end{aligned}
$$



The SEIR matrix decomposition

We can decompose the SEIR model into **F** and **G** rate matrices:

$$\boldsymbol{F}(t) = \begin{bmatrix} 0 & 0 \\ \beta SI & 0 \end{bmatrix}$$



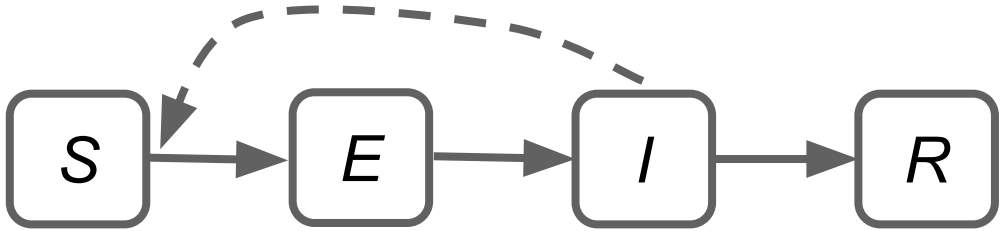
What would the G matrix look like for a SEIR model?

The SEIR matrix decomposition

We can decompose the SEIR model into **F** and **G** rate matrices:

$$\boldsymbol{F}(t) = \begin{bmatrix} 0 & 0 \\ \beta SI & 0 \end{bmatrix}$$

$$\boldsymbol{G}(t) = \begin{bmatrix} 0 & \eta E \\ 0 & 0 \end{bmatrix}$$



Deriving other SIR-type models

The appropriate coalescent model for nearly any type of SIR model can be easily be derived from the Volz structured coalescent.

All we need to do is decompose the model into its component **F** and **G** matrices.

We can then compute the coalescent rates under the model:

$$\lambda_{ij} = \sum_{k}^{m} \sum_{l}^{m} \frac{f_{kl}}{y_k y_l} \left(p_{ik} p_{jl} + p_{il} p_{jk} \right)$$

And use the **G** matrix to track the movement of lineages back through time:

$$\frac{d}{dt} p_{ik} = \sum_{l}^{m} \left(p_{il} g_{kl} - p_{ik} g_{lk} \right)$$

The SEIR model in PhyDyn

PhyDyn is a BEAST2 package for fitting generic SIR models to pathogen phylogenies.

The appropriate coalescent model for a given SIR model is specified by ***F*** and ***G*** matrices .

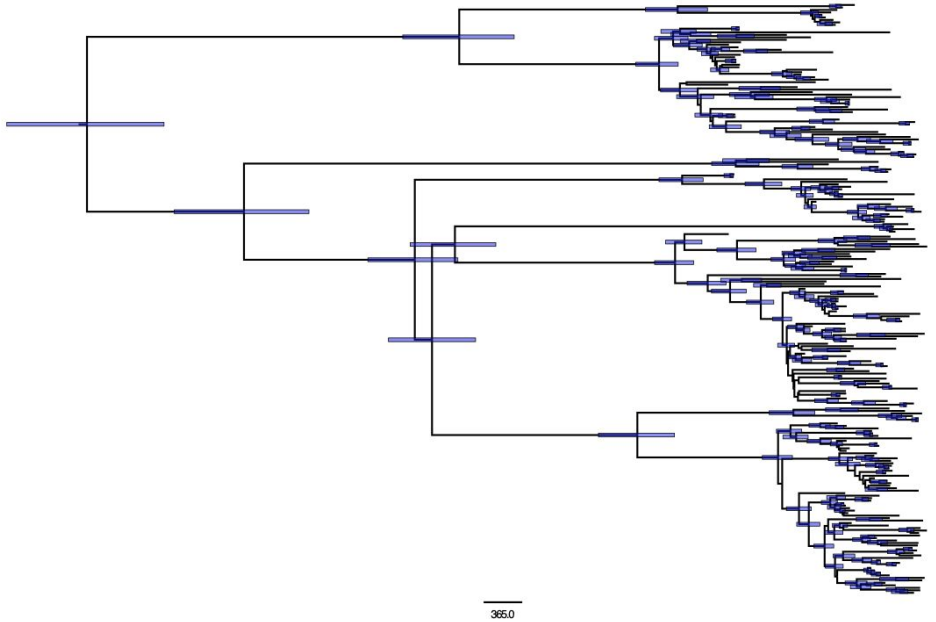
Individual XML elements specify the rates in the ***F*** and ***G*** matrices.

```
<model spec='PopModelODE' id='seirmodel' evaluator="compiled"
     popParams='@initValues' modelParams='@rates'>

    <matrixeq spec='MatrixEquation' type="migration" origin="E" destination="I">
        eta * E
    </matrixeq>

    <matrixeq spec='MatrixEquation' type="birth"  origin="I" destination="E">
        beta * S * I
    </matrixeq>

    <matrixeq spec='MatrixEquation' type="death" origin="Il">
        gamma * I
    </matrixeq>

    <matrixeq spec='MatrixEquation' type="nondeme" origin="R">
        gamma * I
    </matrixeq>
</model>
```

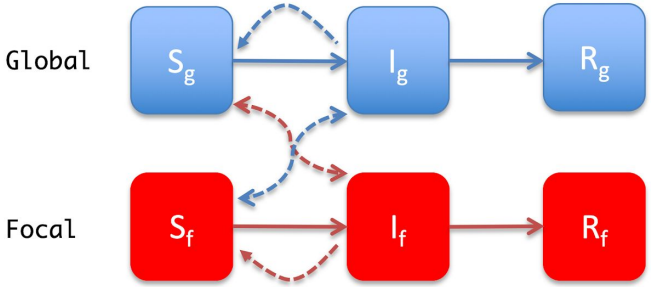
Volz & Siveroni (PLoS Comp Bio, 2018)

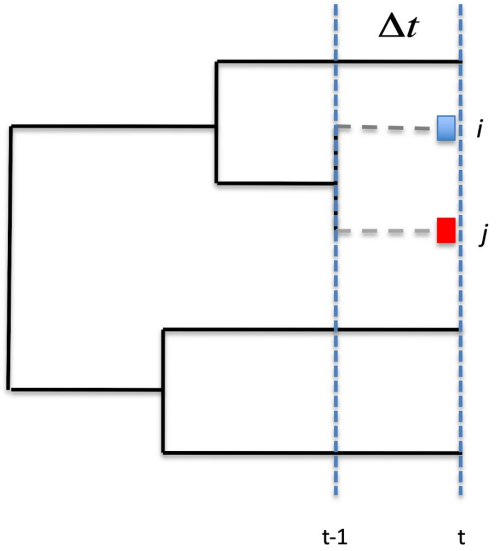
What makes these models powerful is that we can now fit epidemiological models directly to pathogen phylogenies.

Dengue in southern Vietnam



Spatial SIR model

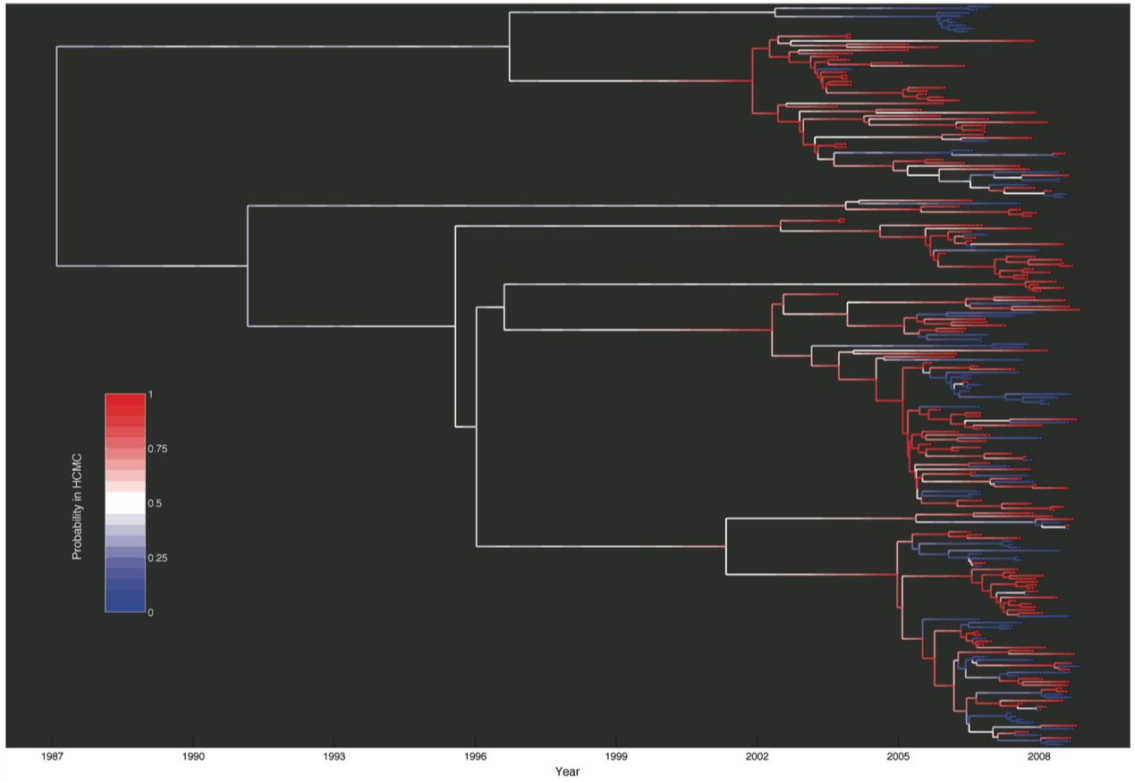




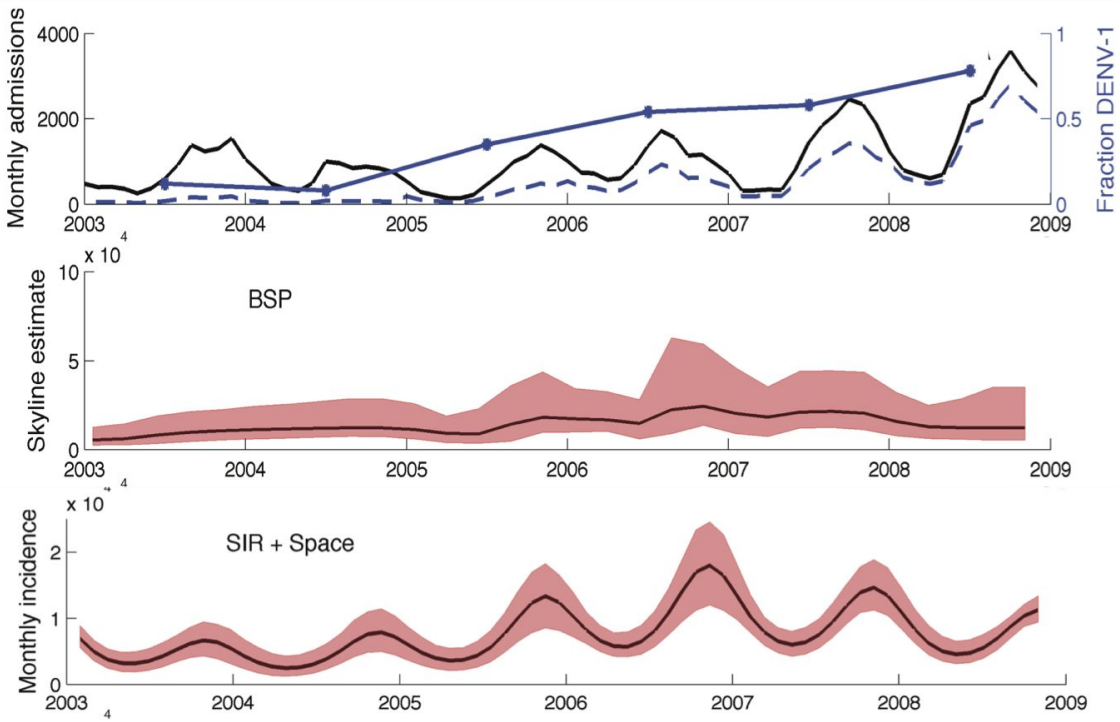
Structured coalescent model:

$$\lambda_{ij} = \sum_{k}^{m} \sum_{l}^{m} \frac{\beta_{kl} \frac{S_l}{N_l} I_k}{I_k I_l} \left(p_{ik} p_{jl} + p_{il} p_{jk} \right)$$

Movement of lineages



Estimates accounting for spatial structure



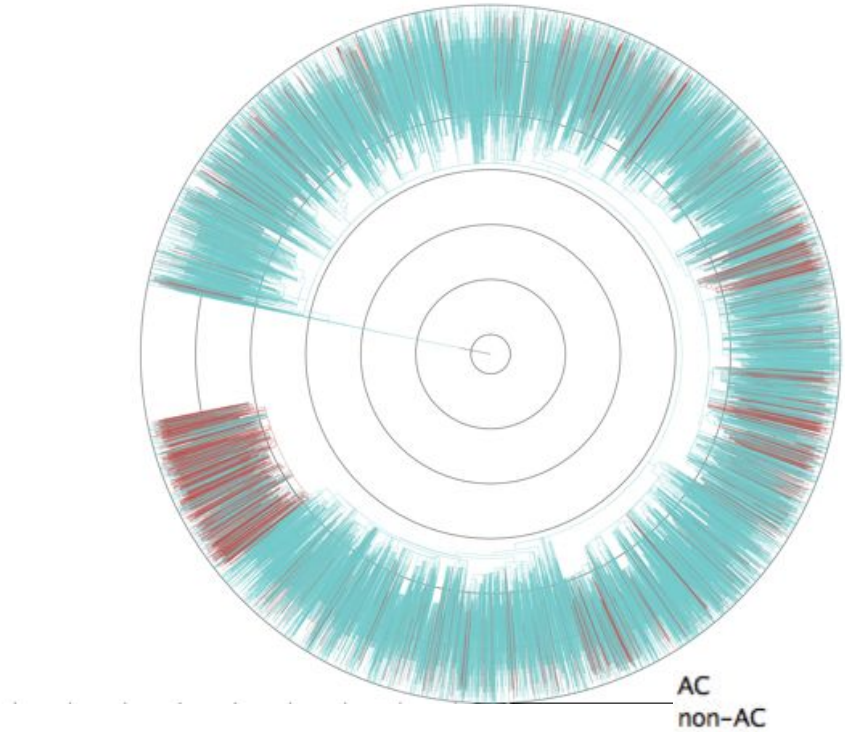
Other uses of phylodynamic models

Reconstructing how disease prevalence and incidence change over time in different populations.

Estimating key epidemiological parameters like R_0 and transmission rates within and between populations.

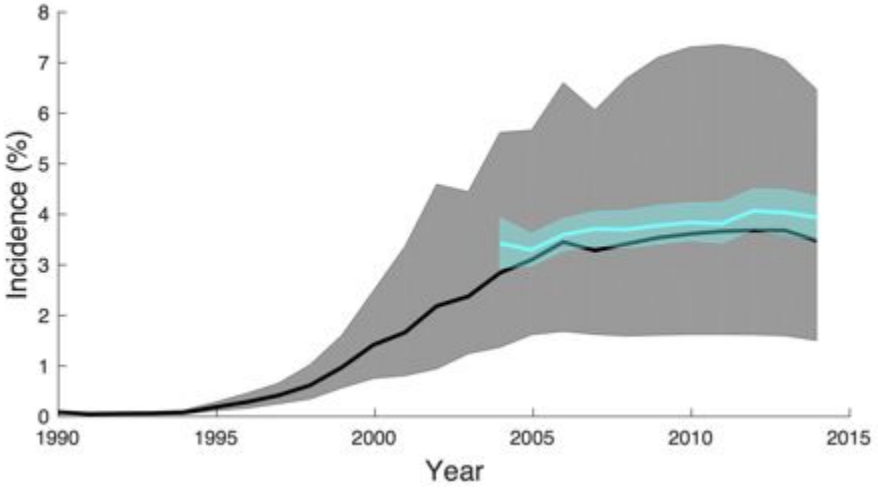
Inferring the sources of transmission including the proportion of infections attributable to a given subpopulation.

HIV in rural Kwa-Zulu Natal



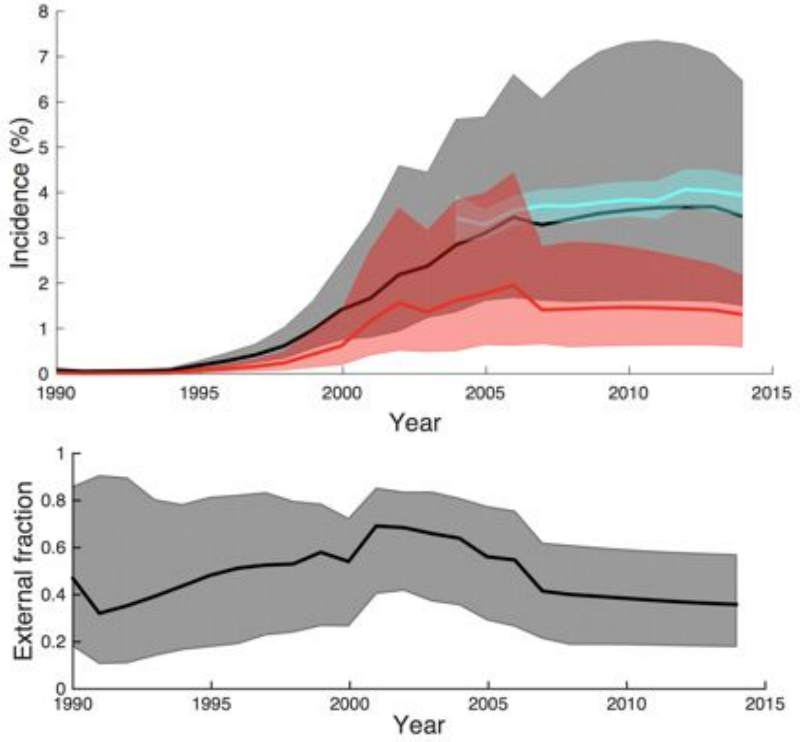
Phylodynamic estimates of HIV incidence

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



Rasmussen et al. (Virus Evolution, 2018)

Incidence due to external introductions



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

Rasmussen et al. (Virus Evolution, 2018)

The PhyDyn tutorial for this week

On Wednesday, we will use the PhyDyn package to fit our own SIR-type models to pathogen phylogenies in BEAST2

First we will test our implementation of the SEI2R model in BEAST using the sequence data we simulated last week.

Then apply to the SEI2R model to some SARS-CoV-2 sequences to estimate key epidemiological parameters for Covid-19.