NC STATE UNIVERSITY

Inferring transmission trees and who’s infecting whom

Molecular Epidemiology of Infectious Diseases

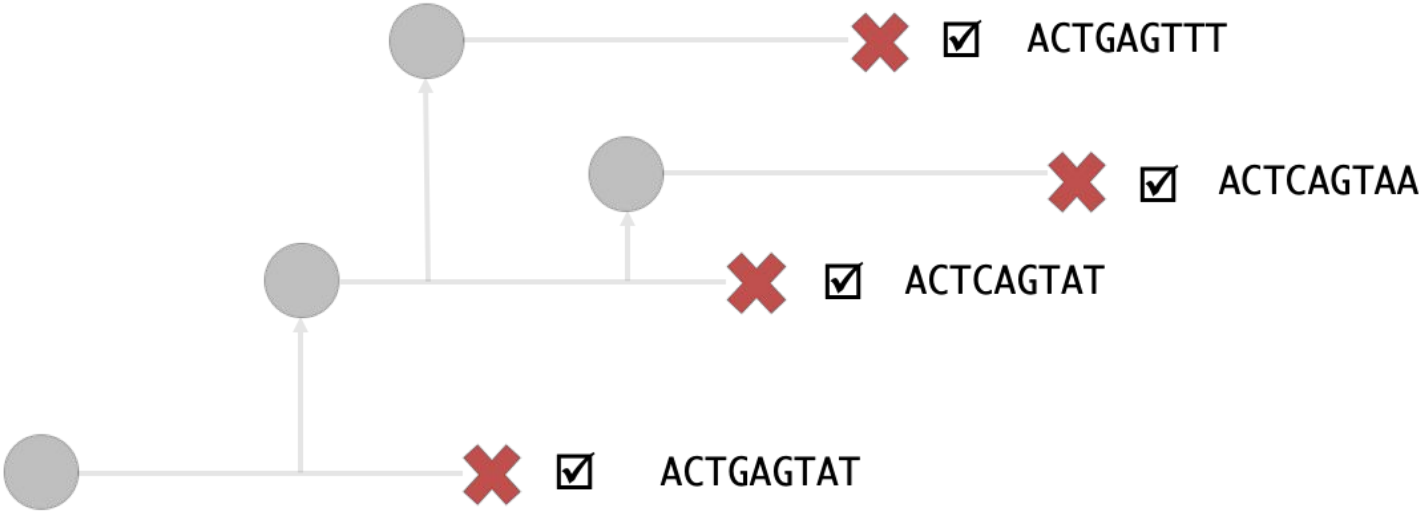
Lecture 5

February 12th, 2024

So far we’ve been focusing on population-level transmission dynamics

Now we will turn to tracking outbreaks at the individual host-scale

A simple epidemic example



Who’s infecting whom

Reconstructing who infected whom is often considered to be the “*holy grail*” of infectious disease epidemiology.

- Identifies who is actually transmitting (e.g. superspreaders)
- Identifies the characteristics of transmitters (e.g. injection drug users)
- Provides a target for control efforts and interventions
- Allows for contact-tracing to prevent further spread

Who’s infecting whom

The unit of infection does not necessarily need to be individual hosts. Transmission tree methods can reconstruct spread among:

- Schools
- Villages
- Fields
- Farms

Two main approaches

1. Methods that directly estimate the underlying transmission tree

2. Methods that reconstruct pathogen phylogenies and then infer transmission routes between hosts

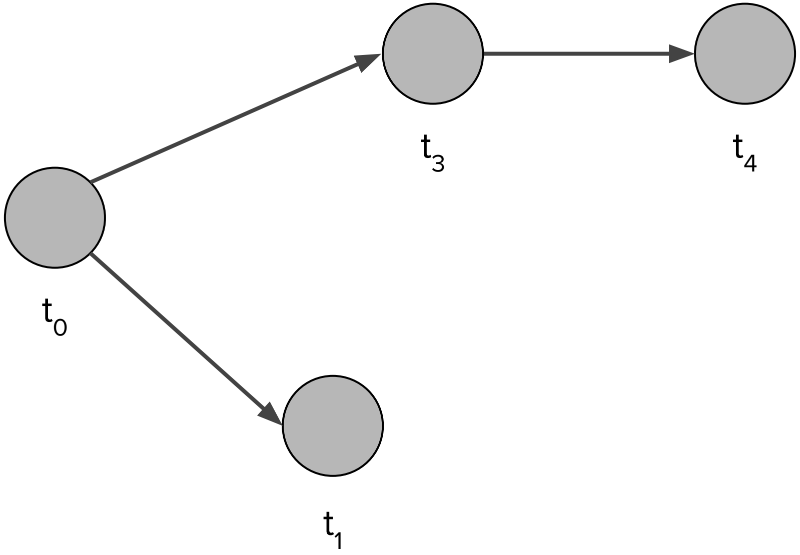
Two main approaches

1. Methods that directly estimate the underlying transmission tree

2. Methods that reconstruct pathogen phylogenies and then infer transmission routes between hosts

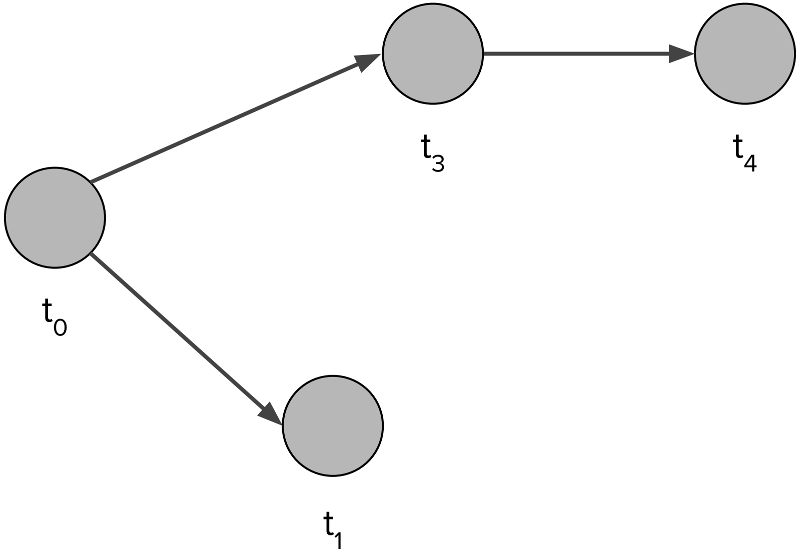
Transmission tree reconstruction

General goal is to probabilistically reconstruct likely transmission chains or links



Transmission tree reconstruction

We often have data on the infection times t_1, t_2, ... t_n and sequences s_1, s_2, ... s_n sampled from each host.



Transmission tree reconstruction

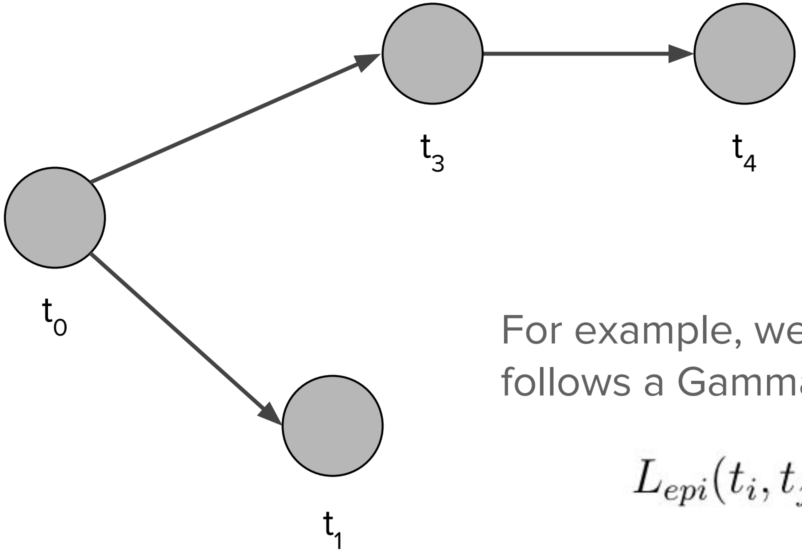
We can divide the problem by thinking about the likelihood of two types of data given a proposed transmission tree:

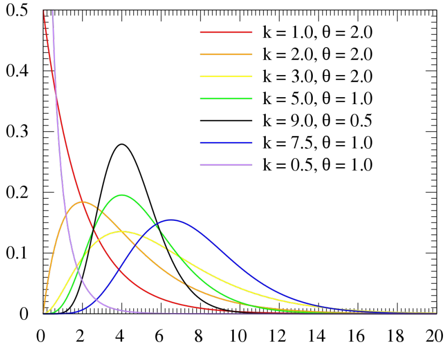
1. The **epidemiological likelihood** of the infection times and any other spatial/temporal data we know about the infected hosts

2. The **genetic likelihood** of the sequence data given a proposed transmission tree

Example: The epidemiological likelihood

The likelihood of a host infected at time t_i infecting another host at time t_j follows a generation time (serial interval) distribution:



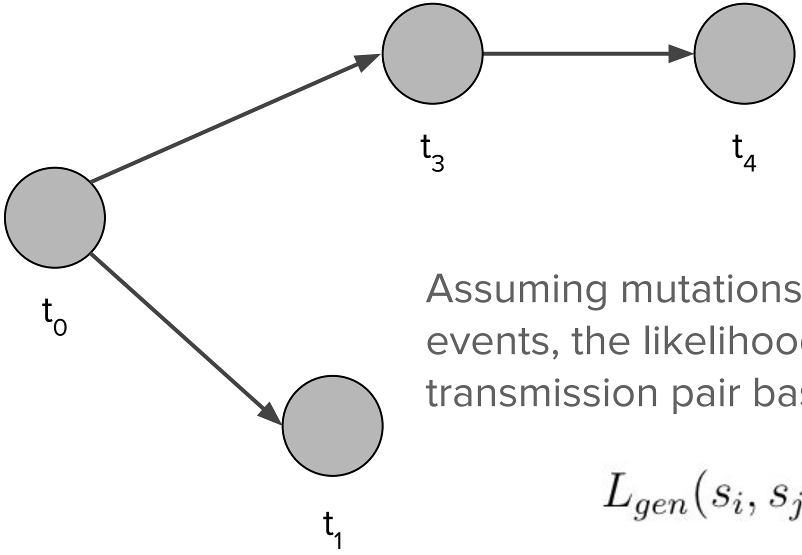


For example, we could assume the generation time follows a Gamma distribution:

$$L_{epi}(t_i, t_j) = Gamma(t_j - t_j|\alpha, \beta)$$

The genetic likelihood (simplest case)

The likelihood of sequences s_i and s_j resulting from a direct transmission between hosts *i* and *j* can be computed based on their genetic distances:



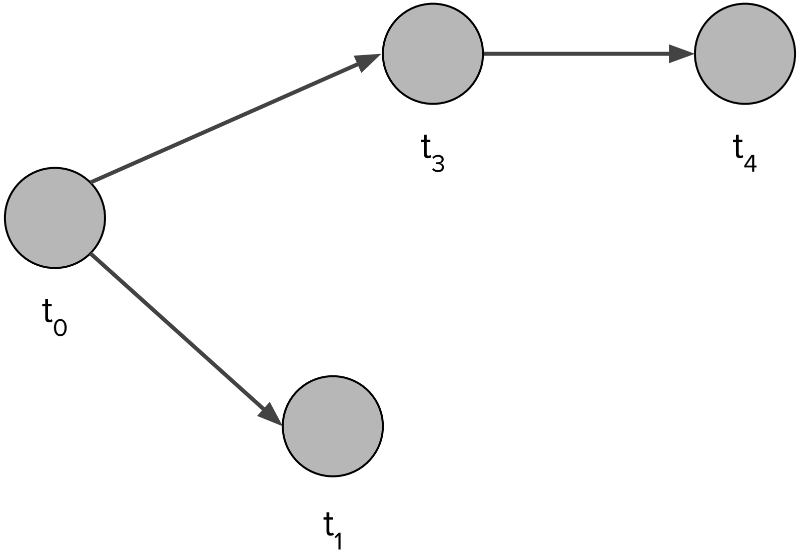
Assuming mutations happen at rate μ at transmission events, the likelihood of two sequences being in a transmission pair based on their genetic distance *d*(s_i,s_j):

$$L_{gen}(s_i, s_j) = \mu^{d(s_i,s_j)}(1-\mu)^{L-d(s_i,s_j)}$$

Jombart *et al.* (2014)

Transmission tree reconstruction

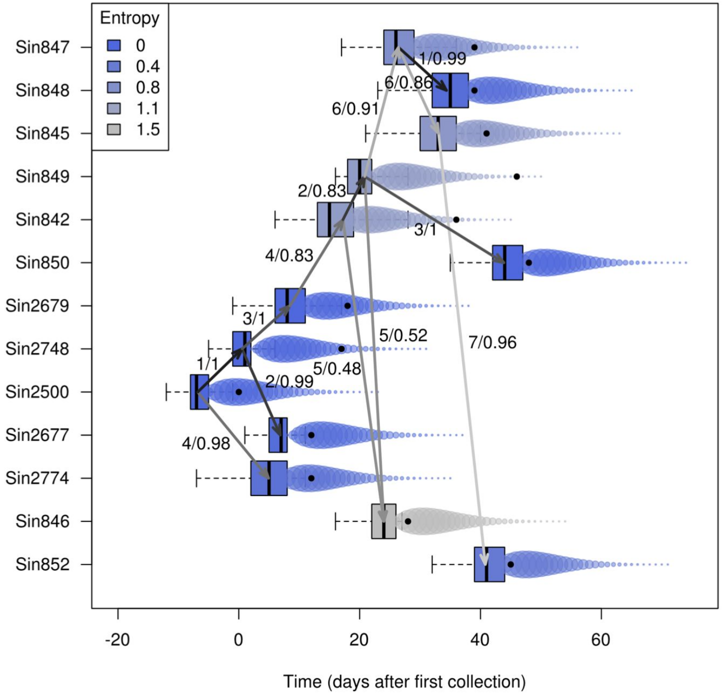
Our goal is to find the transmission tree that maximizes the **overall likelihood** of the infection times and sequence data across all transmission pairs:

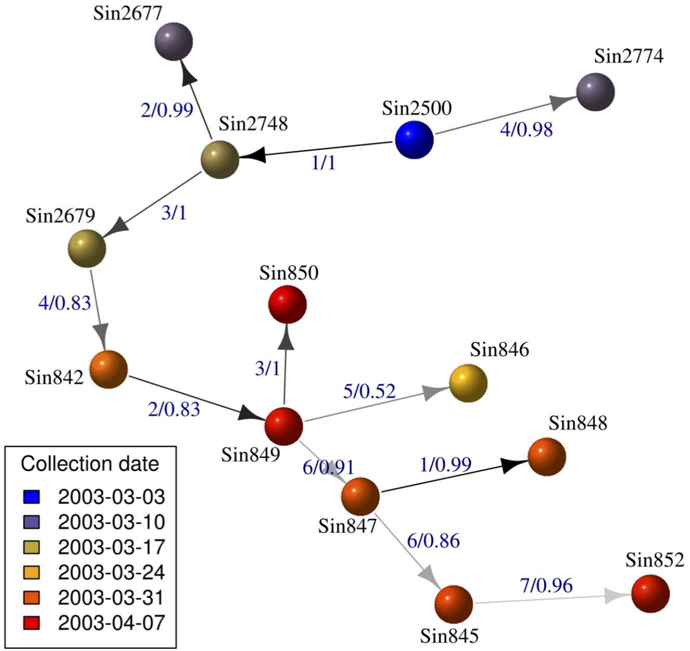


The overall likelihood can be computed as a product over all transmission pairs:

$$L(\mathcal{T}) = \prod_{i,j \in \mathcal{T}} L_{epi}(t_i, t_j) L_{gen}(s_i, s_j)$$

SARS outbreak in Singapore





Jombart *et al.* (2014)

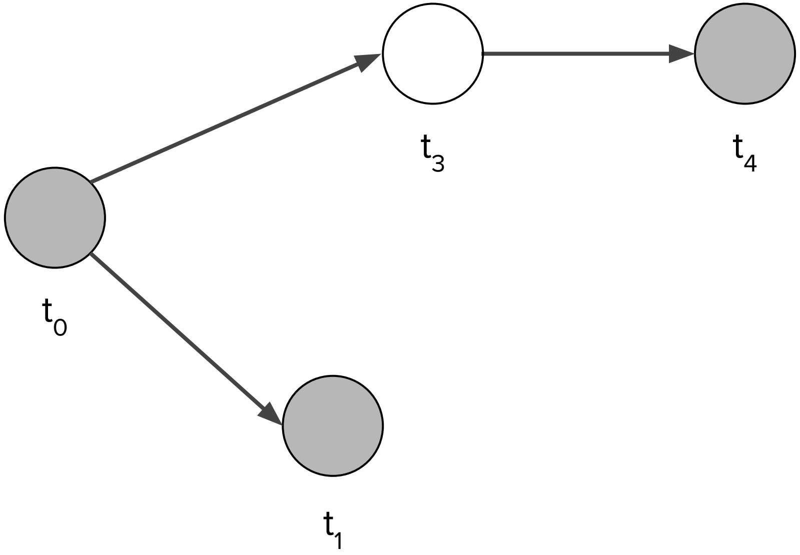
Direct transmission tree reconstruction

Direct reconstruction generally works well when:

- Outbreaks are small and we can sample nearly all infected hosts
- Short and regular generation times
- High between-host genetic divergence but negligible within-host variation

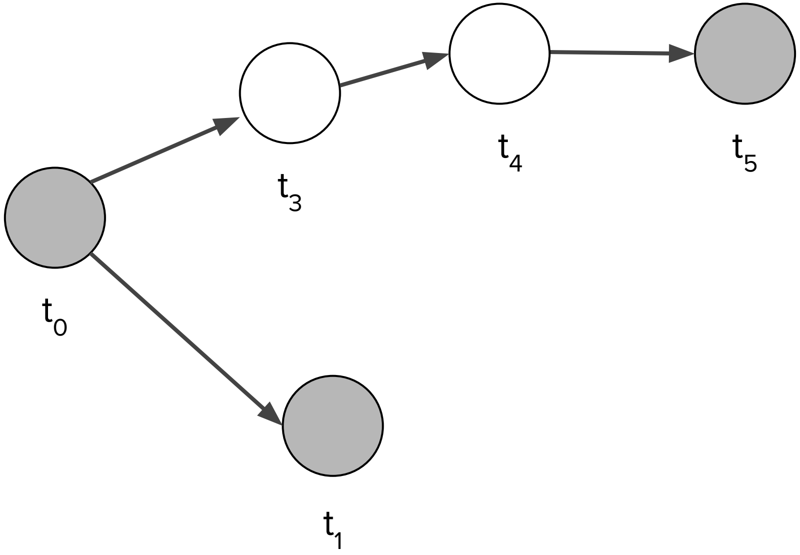
Transmission tree reconstruction

The problem is that we generally have incomplete sampling with at least some unobserved infections.



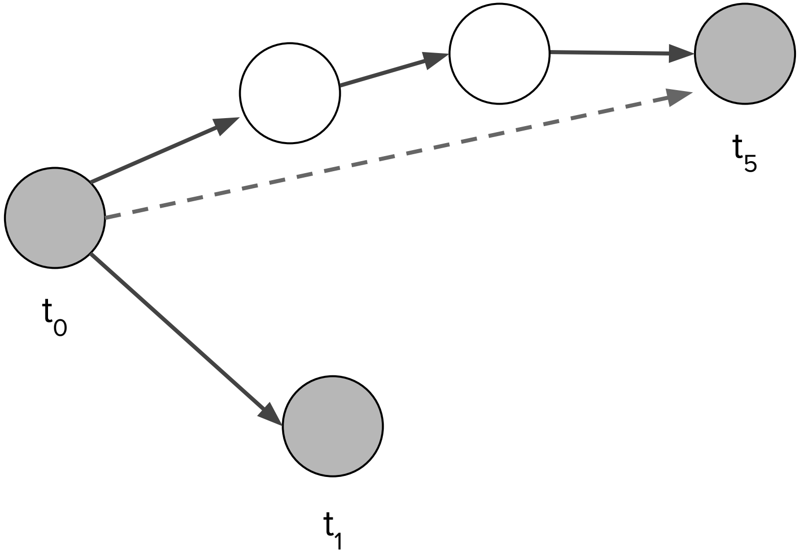
Transmission tree reconstruction

The problem is that we generally have incomplete sampling with at least some unobserved infections.



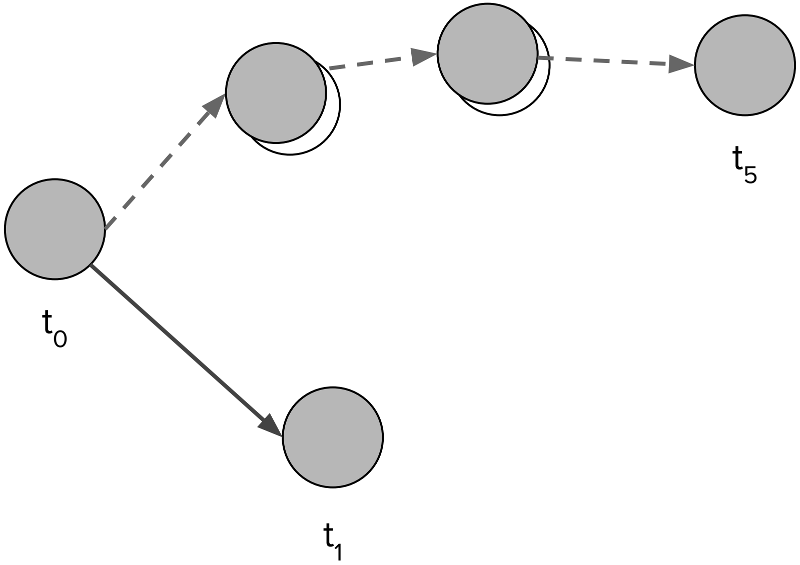
Transmission tree reconstruction

We are therefore likely to misattribute sources of infection to sampled individuals while ignoring unobserved hosts.



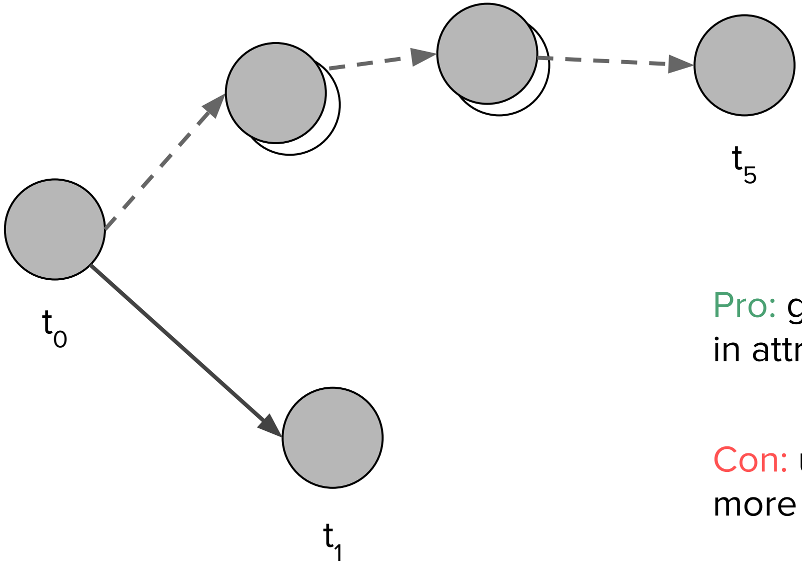
Data augmentation

We can postulate the presence of unobserved infections and impute their presence/absence as additional *latent* (unobserved) variables in the model.



Data augmentation

We can postulate the presence of unobserved infections and impute their presence/absence as additional *latent* (unobserved) variables in the model.



Pro: guards against overconfidence in attributing sources.

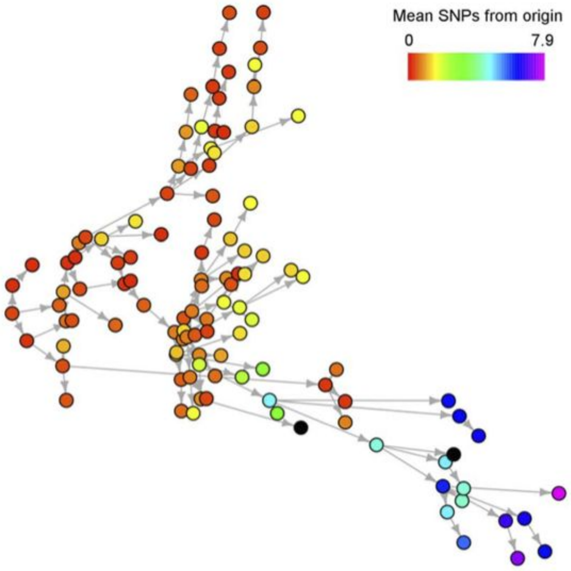
Con: uncertainty will only grow with more unsampled hosts.

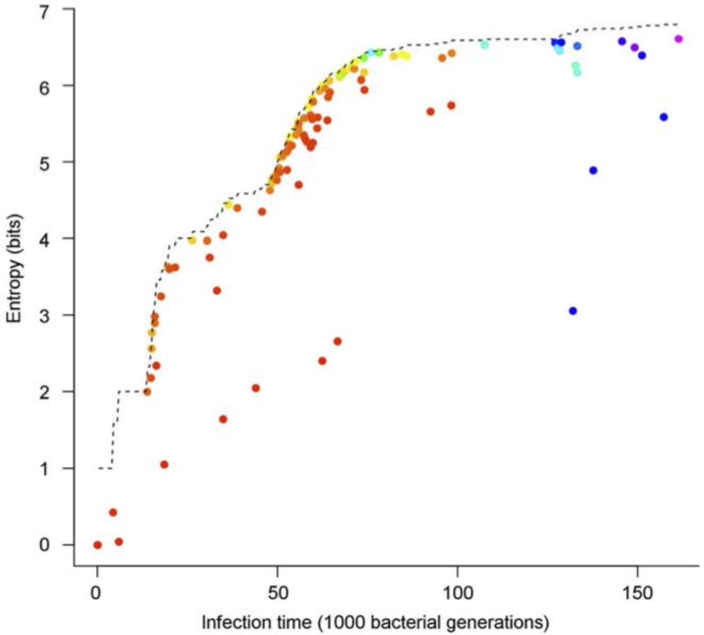
Direct transmission tree reconstruction

Direct reconstruction generally works well when:

- Outbreaks are small and we can sample nearly all infected hosts
- Short and regular generation times
- High between-host genetic divergence but negligible within-host variation

Effect of overlapping infections





Uncertainty

Worby *et al.* (PLoS Comp Bio, 2014)

Direct transmission tree reconstruction

General approach works well when:

- Outbreaks are small and we can sample nearly all infected hosts
- Short and regular generation times
- High between-host genetic divergence but negligible within-host variation

Between-host diversity is often limiting

The number of mutations separating pathogen genomes sampled from direct transmission pairs is often very small (Transmission divergence *<= 1*), providing limited information about who might have infected whom.

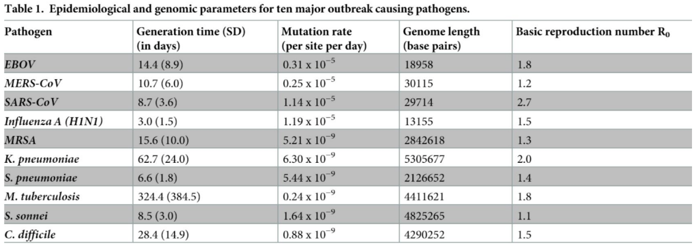
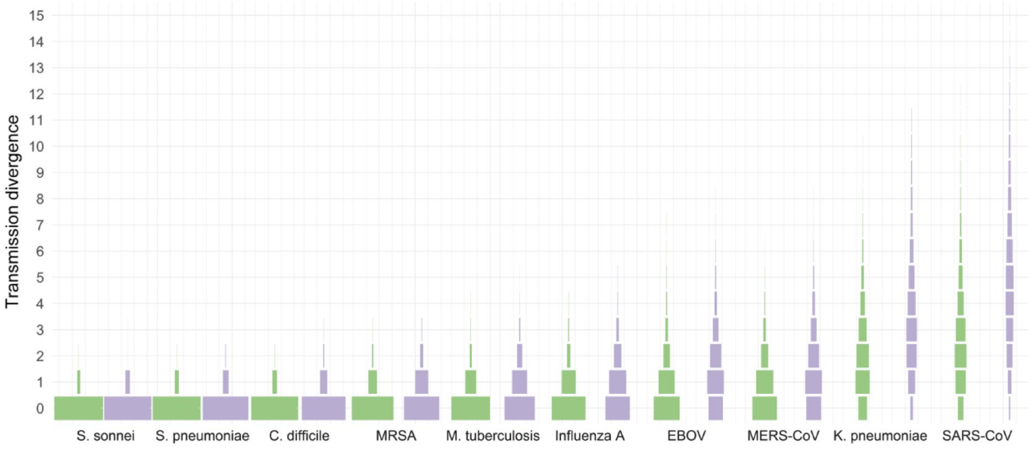


Table 1. Epidemiological and genomic parameters for ten major outbreak causing pathogens.

Pathogen	Generation time (SD) (in days)	Mutation rate (per site per day)	Genome length (base pairs)	Basic reproduction number R_0
EBOV	14.4 (8.9)	0.31×10^{-5}	18958	1.8
MERS-CoV	10.7 (6.0)	0.25×10^{-5}	30115	1.2
SARS-CoV	8.7 (3.6)	1.14×10^{-5}	29714	2.7
Influenza A (H1N1)	3.0 (1.5)	1.19×10^{-5}	13155	1.5
MRSA	15.6 (10.0)	5.21×10^{-9}	2842618	1.3
K. pneumoniae	62.7 (24.0)	6.30×10^{-9}	5305677	2.0
S. pneumoniae	6.6 (1.8)	5.44×10^{-9}	2126652	1.4
M. tuberculosis	324.4 (384.5)	0.24×10^{-9}	4411621	1.8
S. sonnei	8.5 (3.0)	1.64×10^{-9}	4825265	1.1
C. difficile	28.4 (14.9)	0.88×10^{-9}	4290252	1.5



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

Direct transmission tree reconstruction

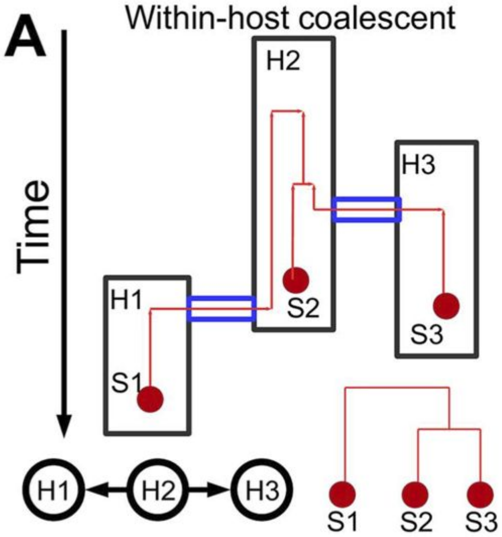
General approach works well when:

- Outbreaks are small and we can sample nearly all infected hosts
- Short and regular generation times
- High between-host genetic divergence but negligible within-host variation

So far we have completely ignored within-host genetic diversity!

Within-host diversity

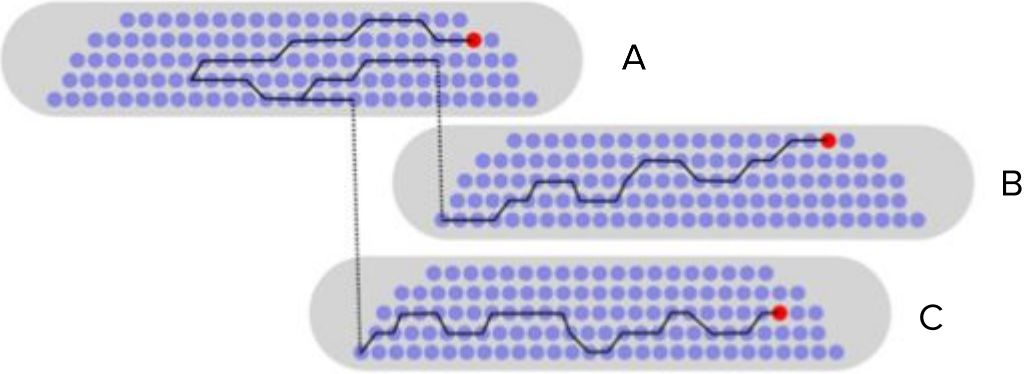
Within-host diversity can cause discordance between pathogen phylogenies and the transmission tree.



De Maio *et al.* (PLoS Comp Bio, 2016)

Within-host diversity

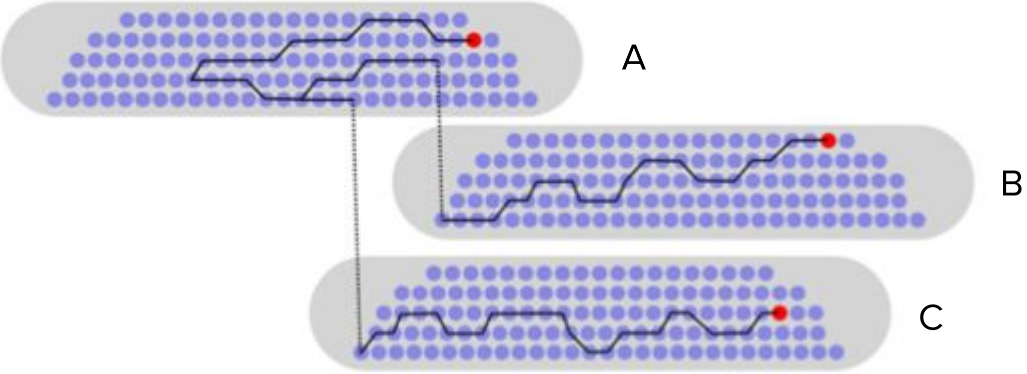
The branching structure of the phylogeny will depend on the timing and order of coalescent events within hosts



Ympa *et al.* (Genetics, 2013)

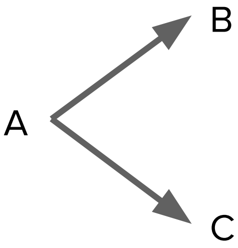
Within-host diversity

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



Ympa *et al.* (Genetics, 2013)

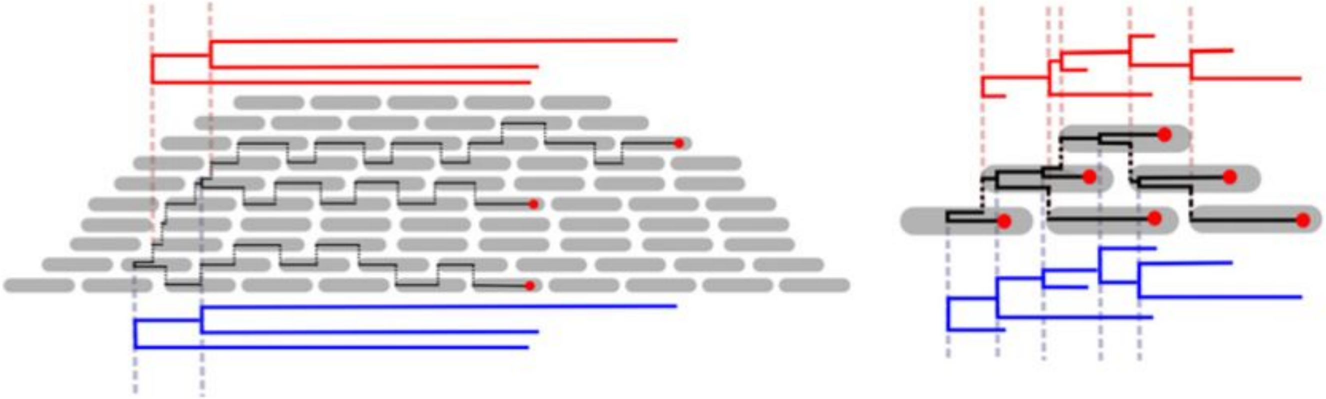
Actual transmission tree:



But three different phylogenetic relationships are possible!

Within-host diversity

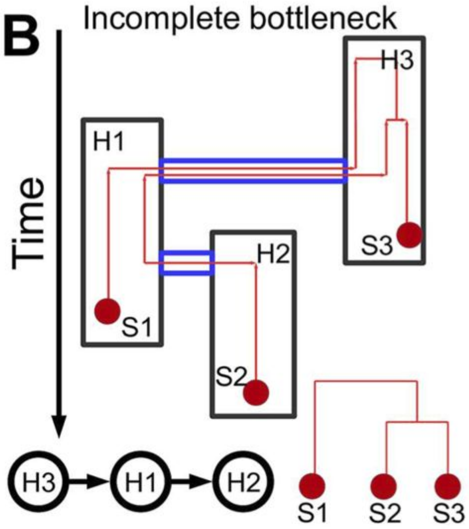
If two lineages coalesce at a transmission event, the coalescent event will always occur before the actual transmission event



Ympa *et al.* (Genetics, 2013)

Within-host diversity

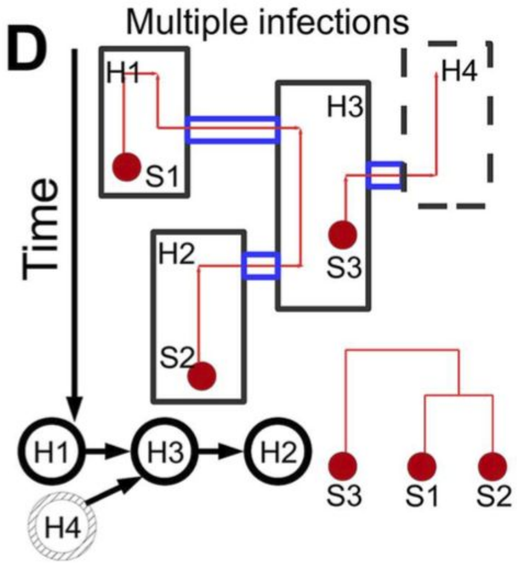
Incomplete transmission bottlenecks can lead to even more extreme discrepancies between transmission trees and phylogenies



De Maio *et al.* (PLoS Comp Bio, 2016)

Within-host diversity

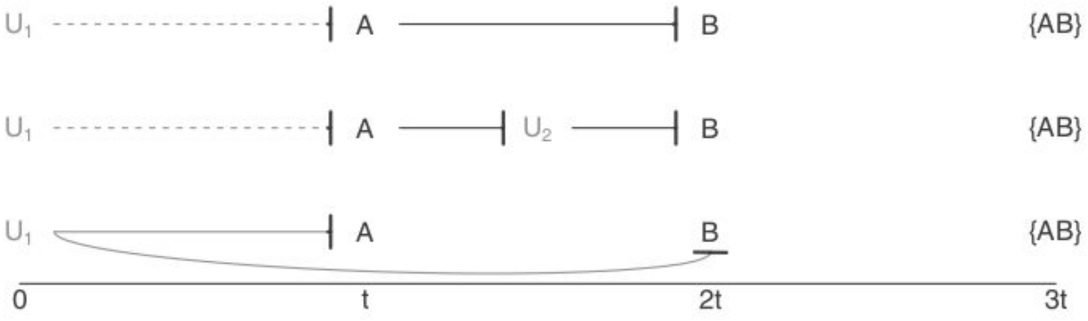
Multiple infections can cause hosts to be erroneously excluded from transmission chains.



De Maio *et al.* (PLoS Comp Bio, 2016)

Within-host diversity

But on the positive side, within-host diversity can also help link infections and resolve the directionality of transmission between a donor and recipient.



Direct transmission

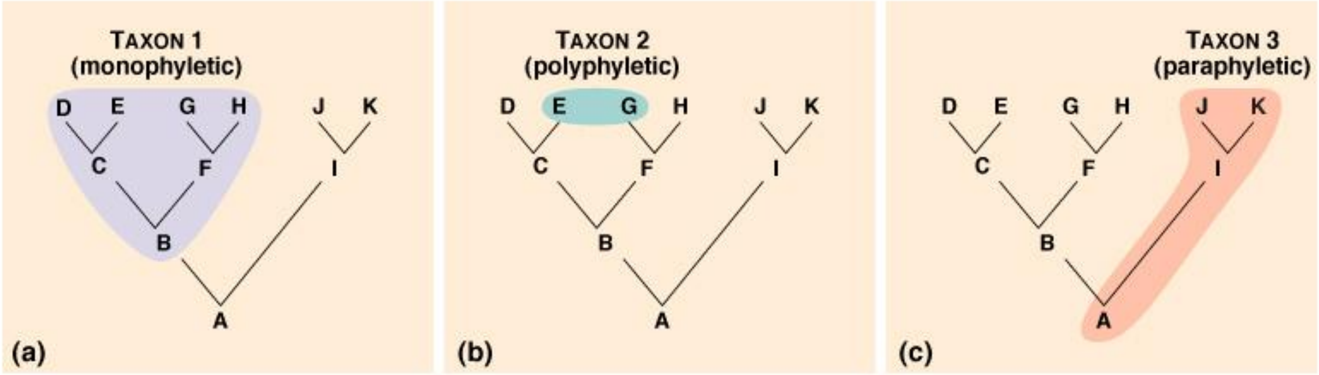
Unsampled intermediate

Common source

Romero-Severson *et al.* (PNAS, 2016)

Phyletic relationships

The **phyletic relationships** among sampled pathogens can provide information about the source of transmission if we have multiple samples from each host.



©1999 Addison Wesley Longman, Inc.

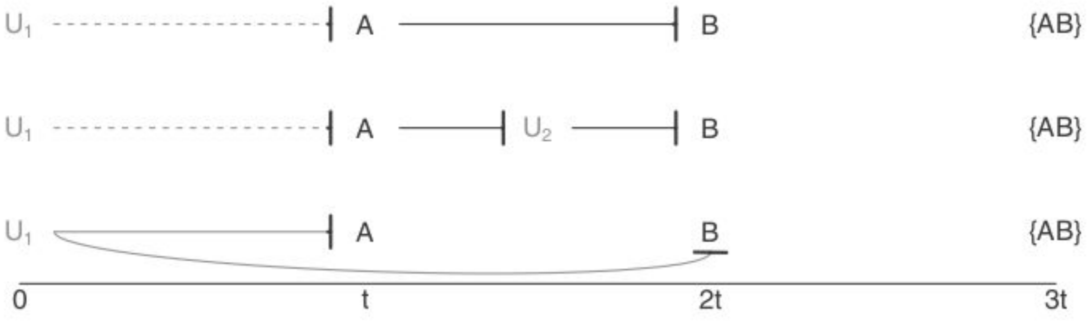
Within-host diversity

Let’s consider the different phyletic relationships among lineages samples from the transmission pair A-B:

Direct transmission

Unsampled intermediate

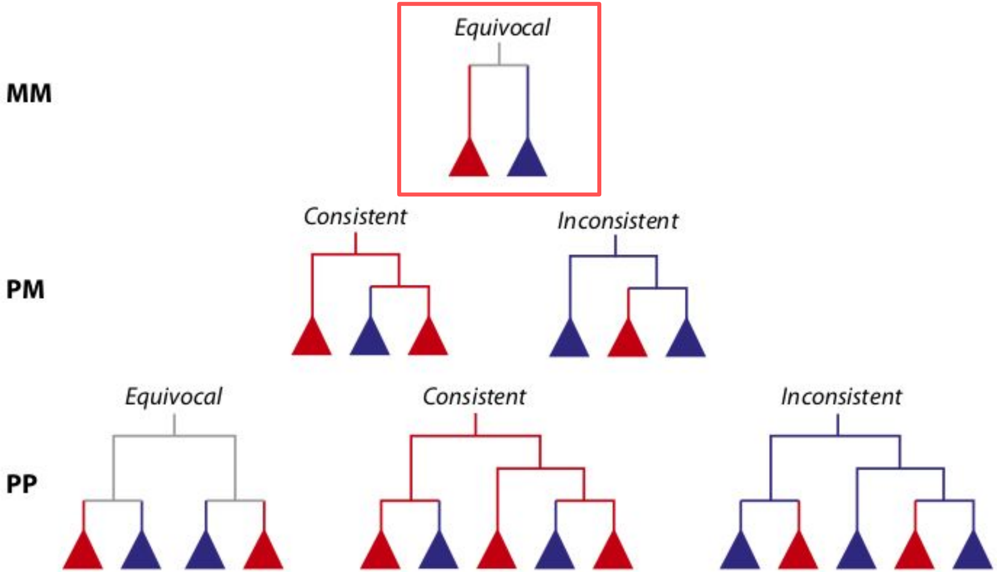
Common source



Romero-Severson *et al.* (PNAS, 2016)

Phyletic relationships

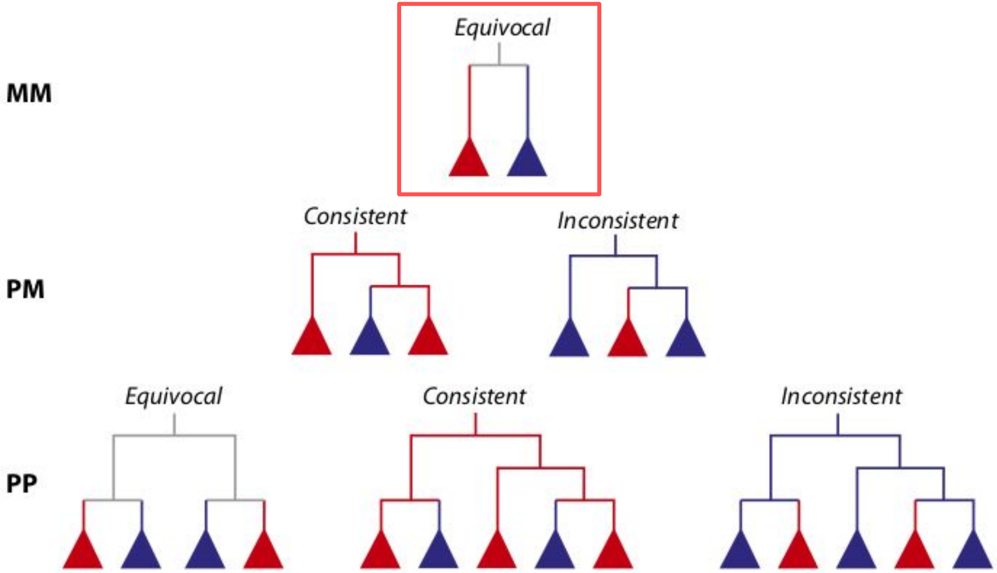
The **phyletic relationships** among sampled lineages can provide information about the source of transmission if we have multiple samples from each host.



Romero-Severson *et al.* (PNAS, 2016)

Phyletic relationships

The **phyletic relationships** among sampled lineages can provide information about the source of transmission if we have multiple samples from each host.

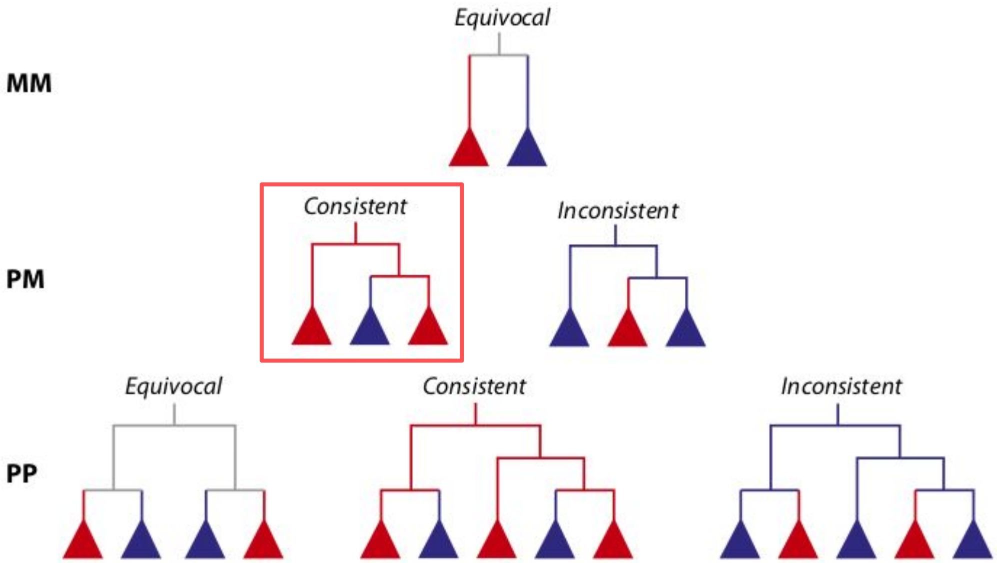


Monophyletic-Monophyletic (MM): Equivocal about the directionality of transmission, but likely to result from a common source of transmission

Romero-Severson *et al.* (PNAS, 2016)

Phyletic relationships

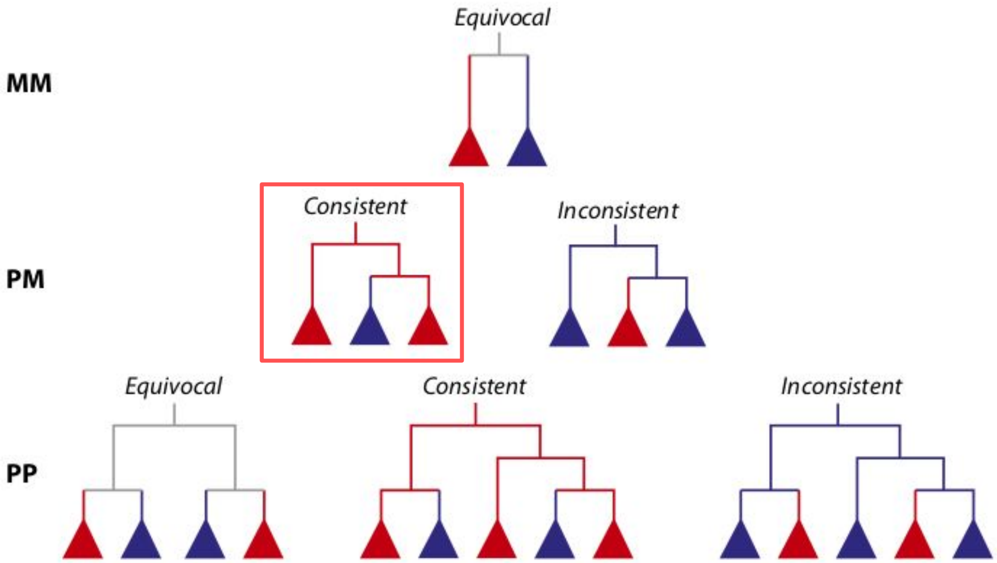
The **phyletic relationships** among sampled lineages can provide information about the source of transmission if we have multiple samples from each host.



Romero-Severson *et al.* (PNAS, 2016)

Phyletic relationships

The **phyletic relationships** among sampled lineages can provide information about the source of transmission if we have multiple samples from each host.

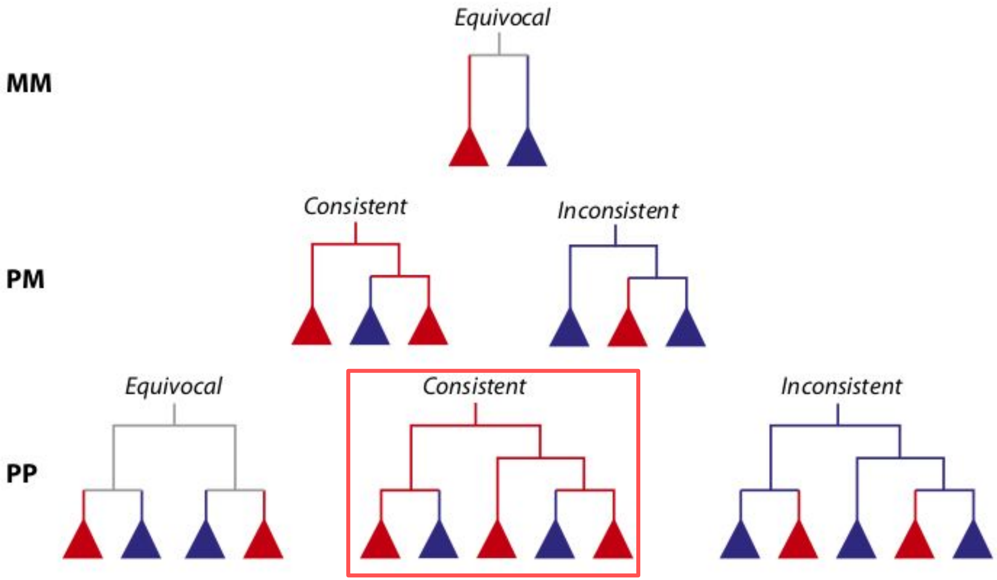


Paraphyletic-Monophyletic (PM): Donor is generally paraphyletic (red) while the recipient (blue) is monophyletic. Most likely results from direct or indirect transmission.

Romero-Severson *et al.* (PNAS, 2016)

Phyletic relationships

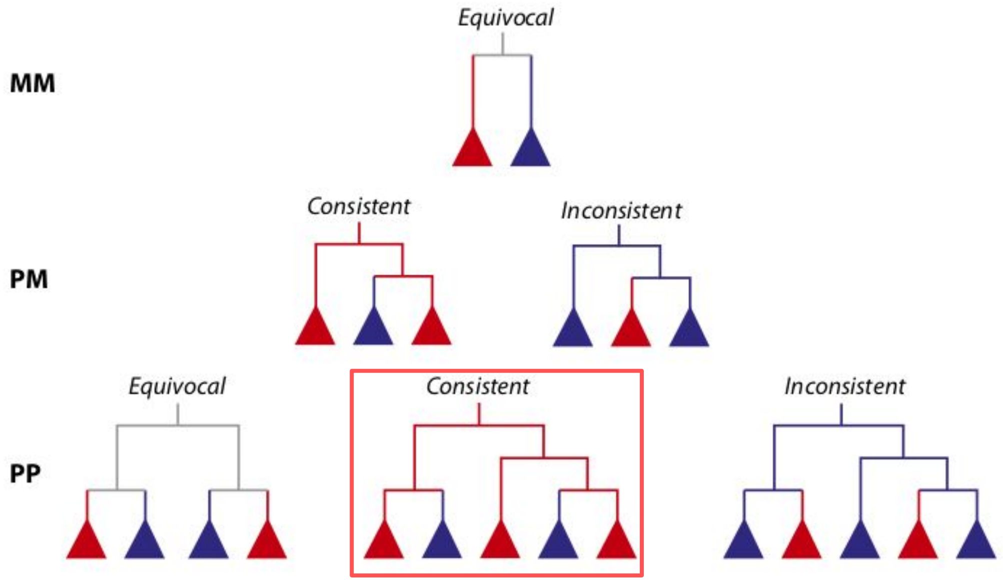
The **phyletic relationships** among sampled lineages can provide information about the source of transmission if we have multiple samples from each host.



Romero-Severson *et al.* (PNAS, 2016)

Phyletic relationships

The **phyletic relationships** among sampled lineages can provide information about the source of transmission if we have multiple samples from each host.



Paraphyletic-Polyphyletic (PP): Generally indicates direct transmission between donor (paraphyletic) and recipient (polyphyletic). Indirect transmission very improbable.

Romero-Severson *et al.* (PNAS, 2016)

Two main approaches

1. Methods that directly estimate the underlying transmission tree

2. Methods that reconstruct pathogen phylogenies and then infer transmission events between hosts

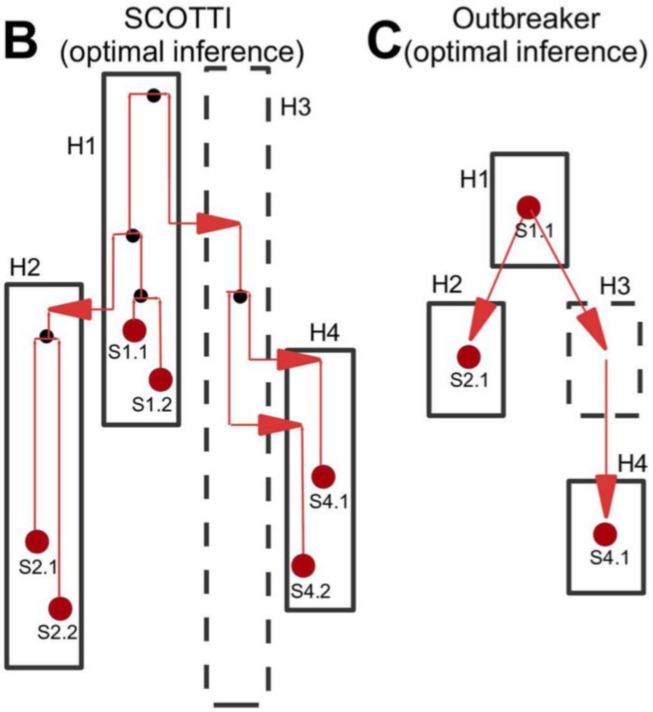
The SCOTTI Approach

Structured COalescent Transmission Tree Inference

Treats each host as a different subpopulation in a structured coalescent model.

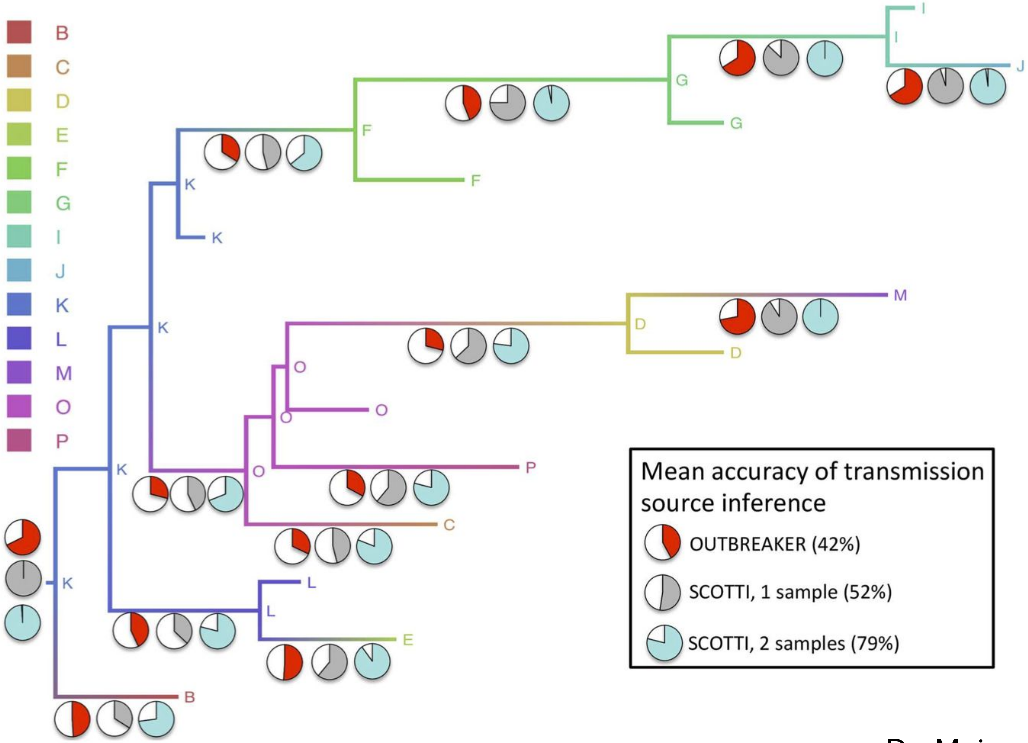
Inferred migration events can be used to reconstruct transmission routes

Accounts for within-host diversity, unsampled hosts and incomplete transmission bottlenecks



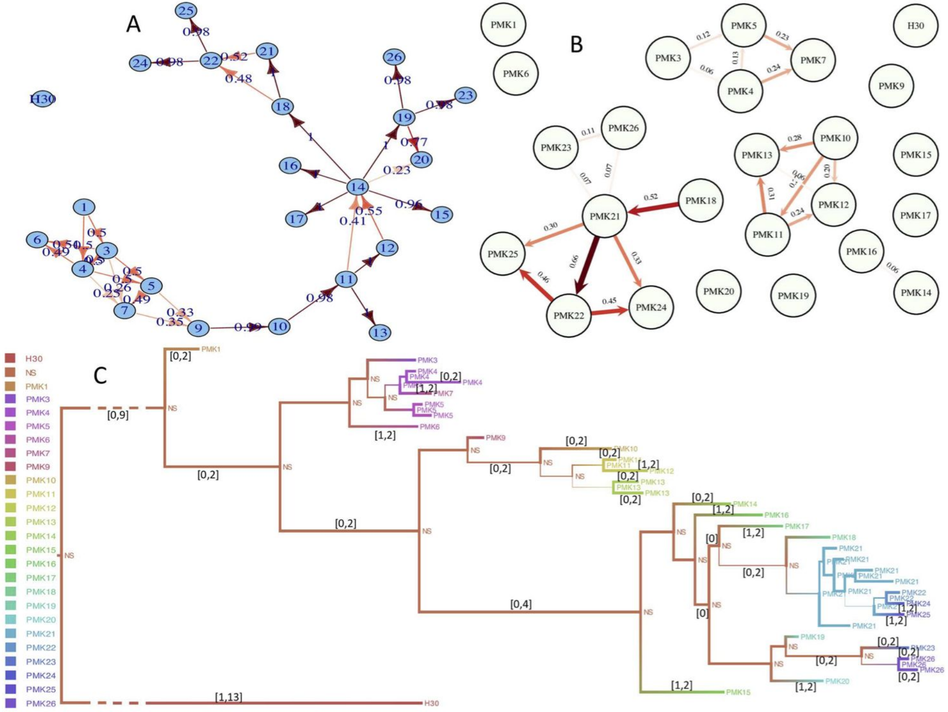
De Maio *et al.* (PLoS Comp Bio, 2016)

SCOTTI versus Outbreaker



De Maio *et al.* (PLoS Comp Bio, 2016)

Klebsiella outbreak reconstruction



De Maio *et al.* (PLoS Comp Bio, 2016)

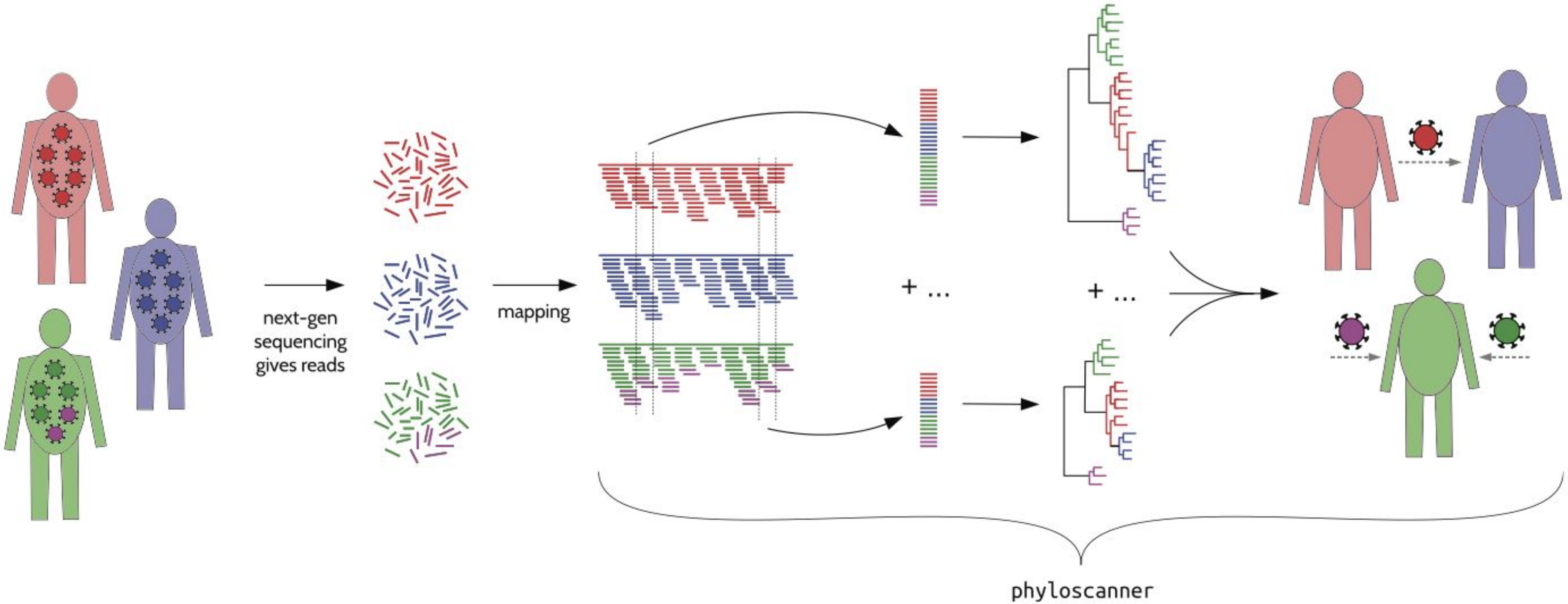
Summary

We can reconstruct transmission trees directly from genetic data or in combination with additional epidemiological data.

Reconstructing transmission trees from genetic data alone is very difficult especially if there are many unsampled hosts and high within-host genetic diversity.

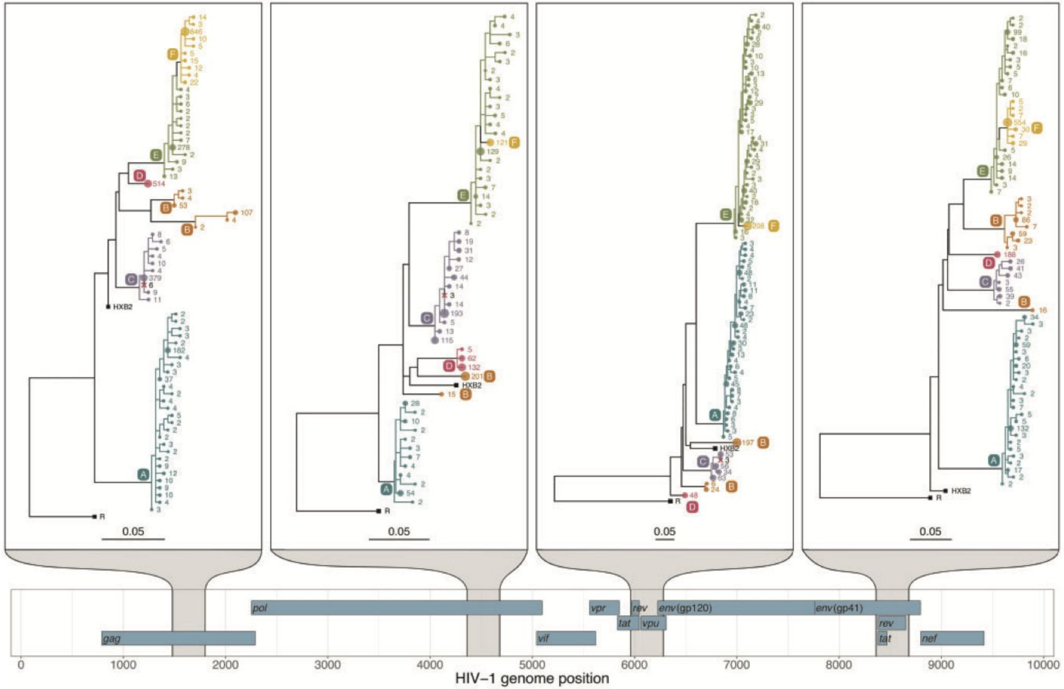
Newer (phylogenetic) approaches leverage the ability to sequence multiple pathogens from each host to more accurately reconstruct transmission chains.

The phyloscanner approach



Wymant *et al.* (MBE, 2017)

The phyloscanner approach



Wymant *et al.* (MBE, 2017)