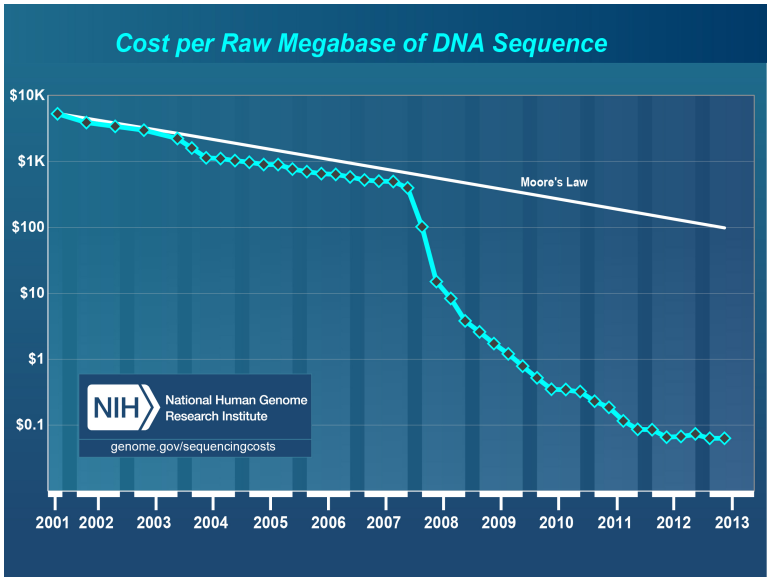
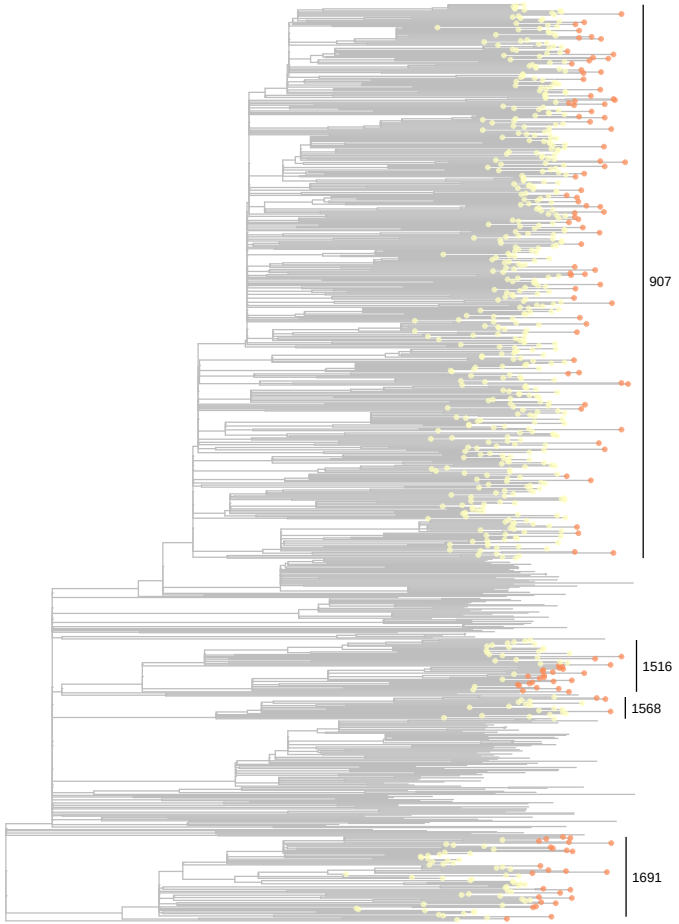
# Genomic epidemiology with TransPhylo: methods, applications and limitations

Caroline Colijn

# Sequencing for pathogens: there is a lot of data



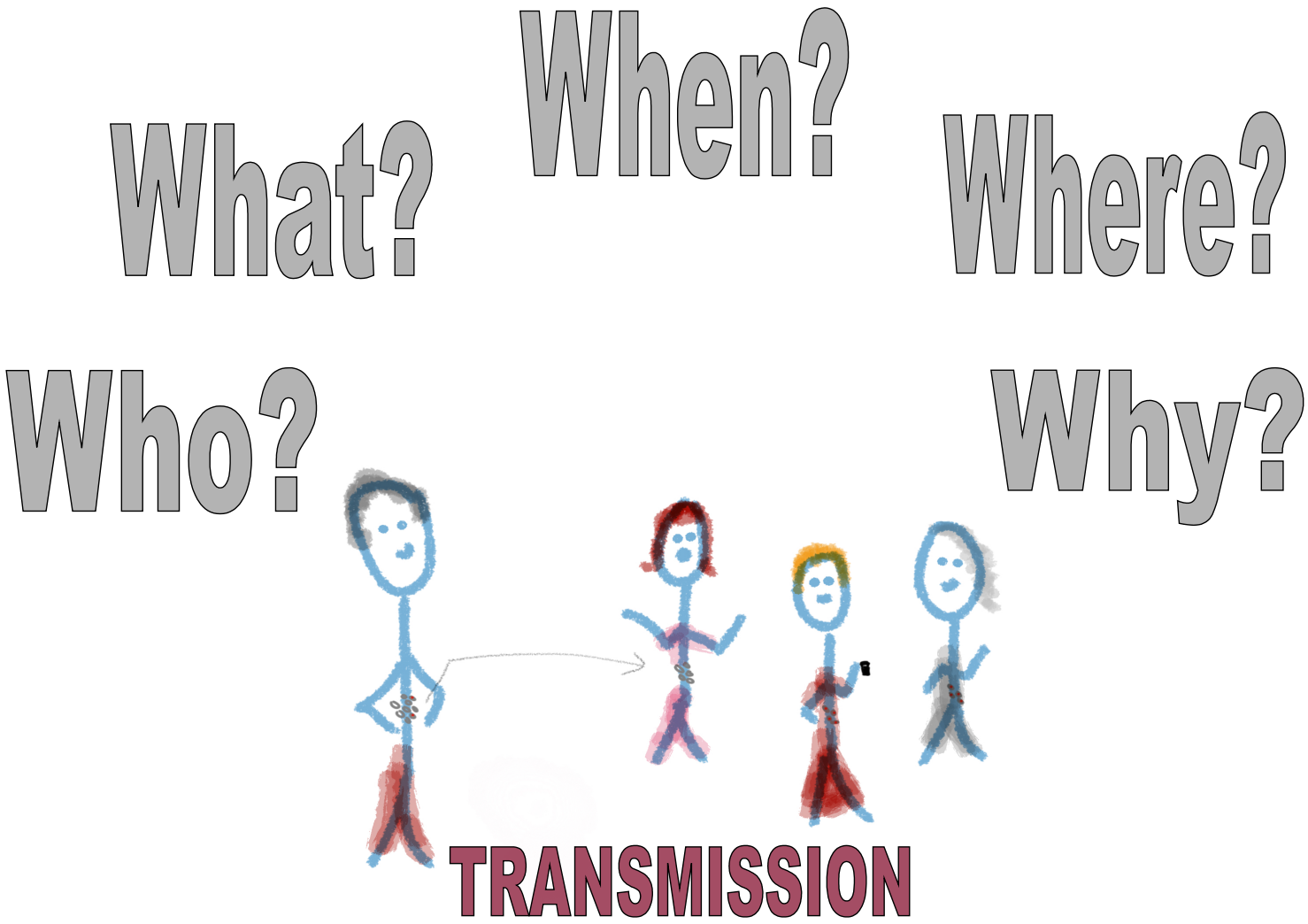

Sequencing is cheap now



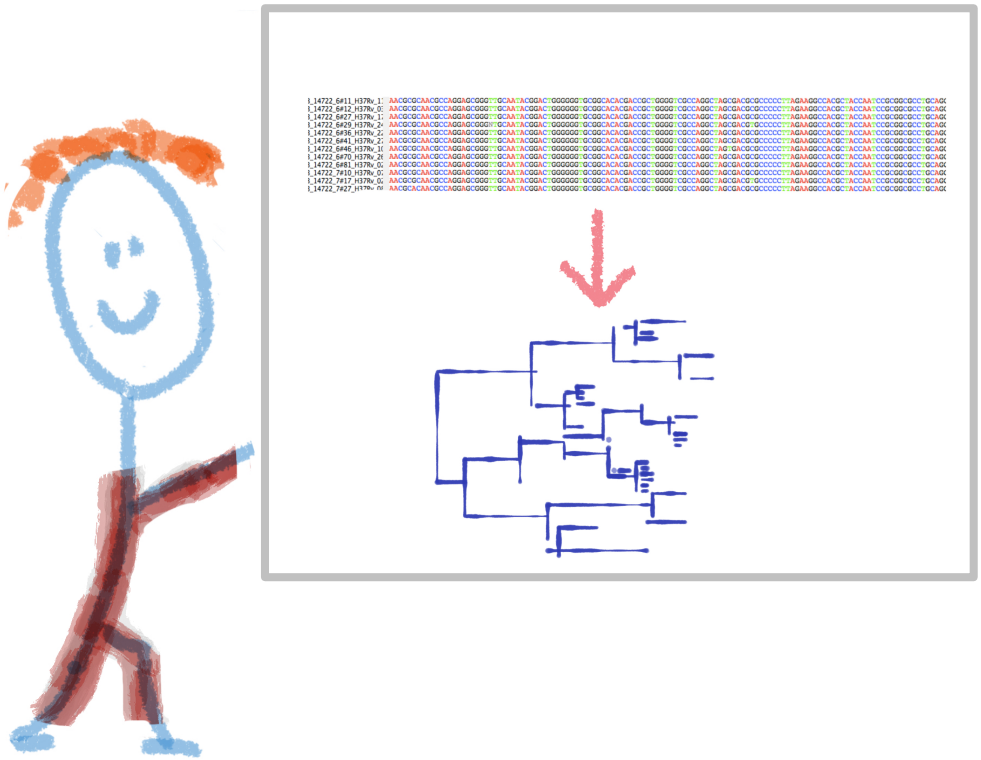

Recent influenza H3N2 subset

We can sequence *lots* of pathogen: thousands per study

# We want to understand transmission



# We have sequences



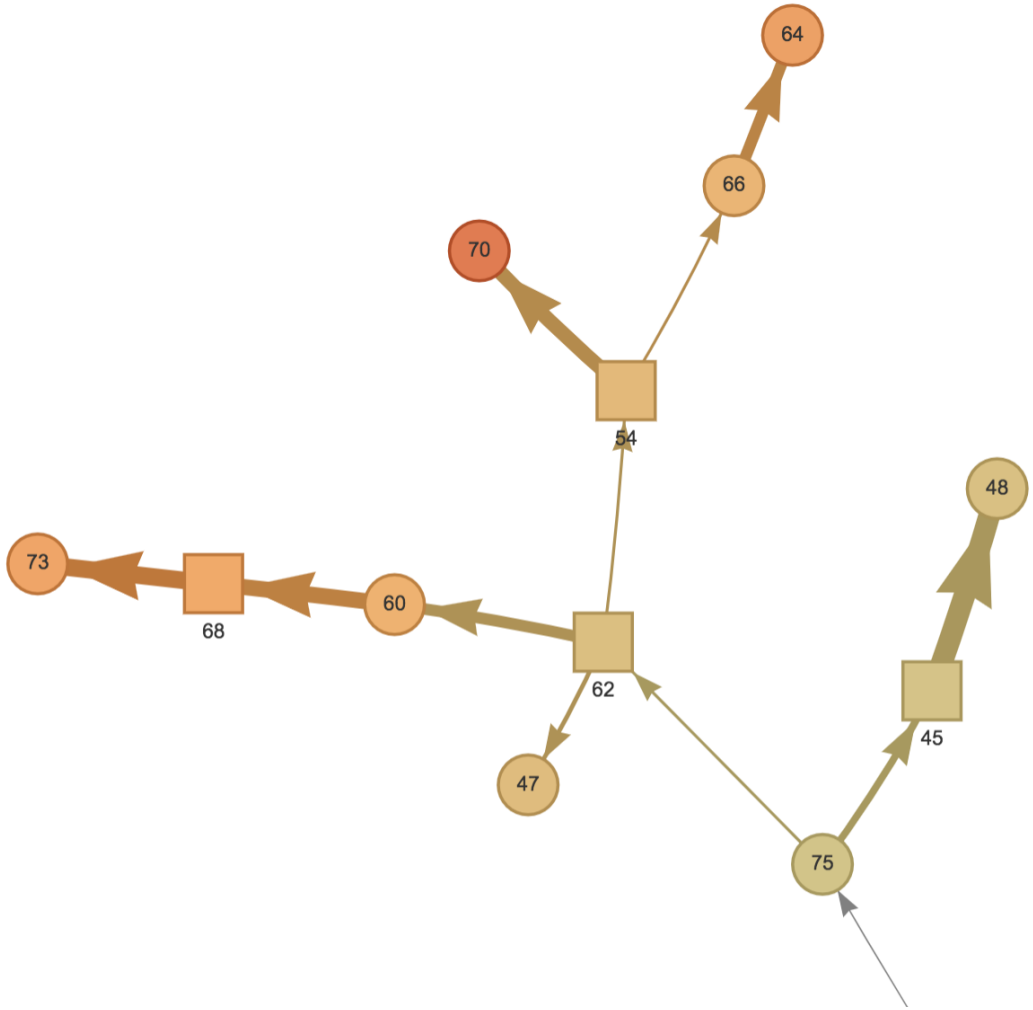
Phylogeny: tips – taxa (sequences). Internal nodes – inferred common ancestors of groups of tips.

# REMEMBER THE OLD GAME OF "TELEPHONE?

# Transmission tree

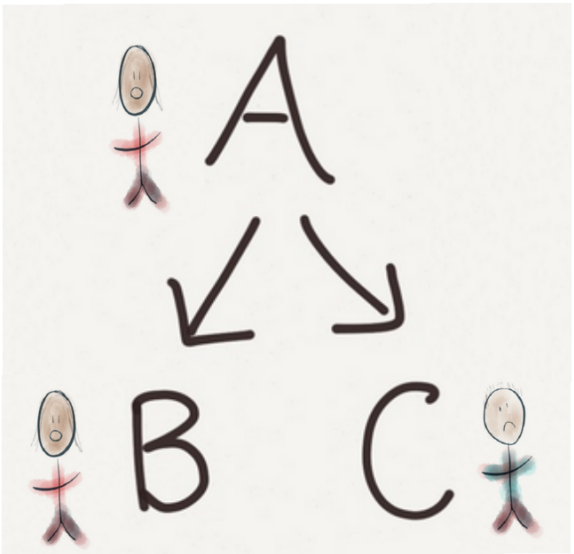
Definition: A *transmission tree* is a tree in which nodes are people and edges (directed) correspond to infection events.

Edges may be associated with times of infection.

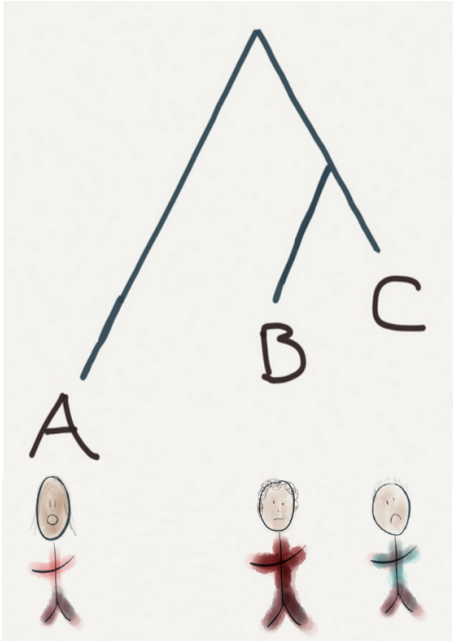



QUESTION: How are transmission trees and phylogenies related?

# Example: transmission tree and phylogeny




A infects B and C




Phylogeny

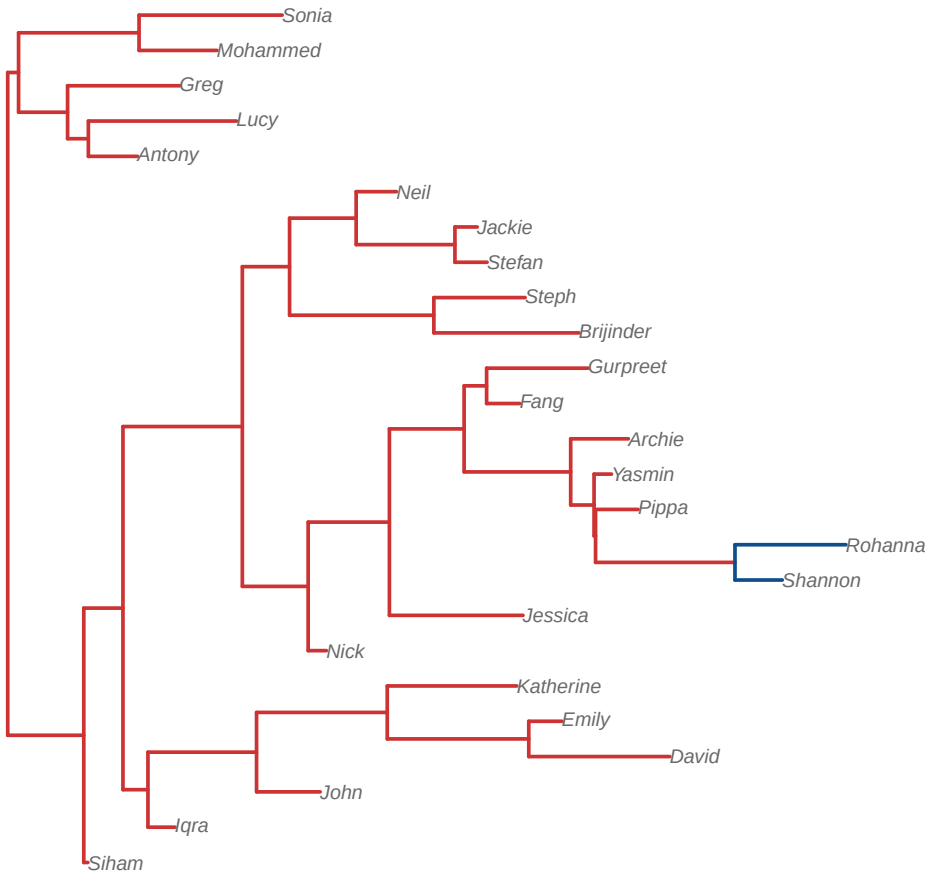
# Transmissions cannot all be phylogenetic pairs

You can only be in a pair with one other.

But you could infect many others.

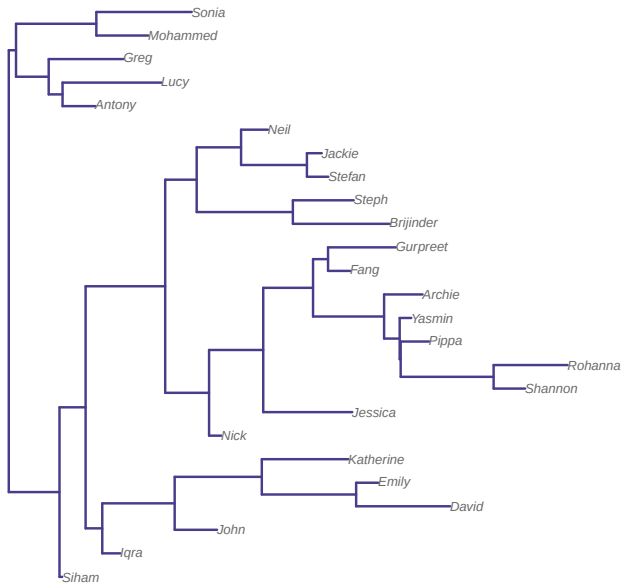
Not all transmission events can be phylo pairs.

Not all closely related pairs of sequences are transmission events.

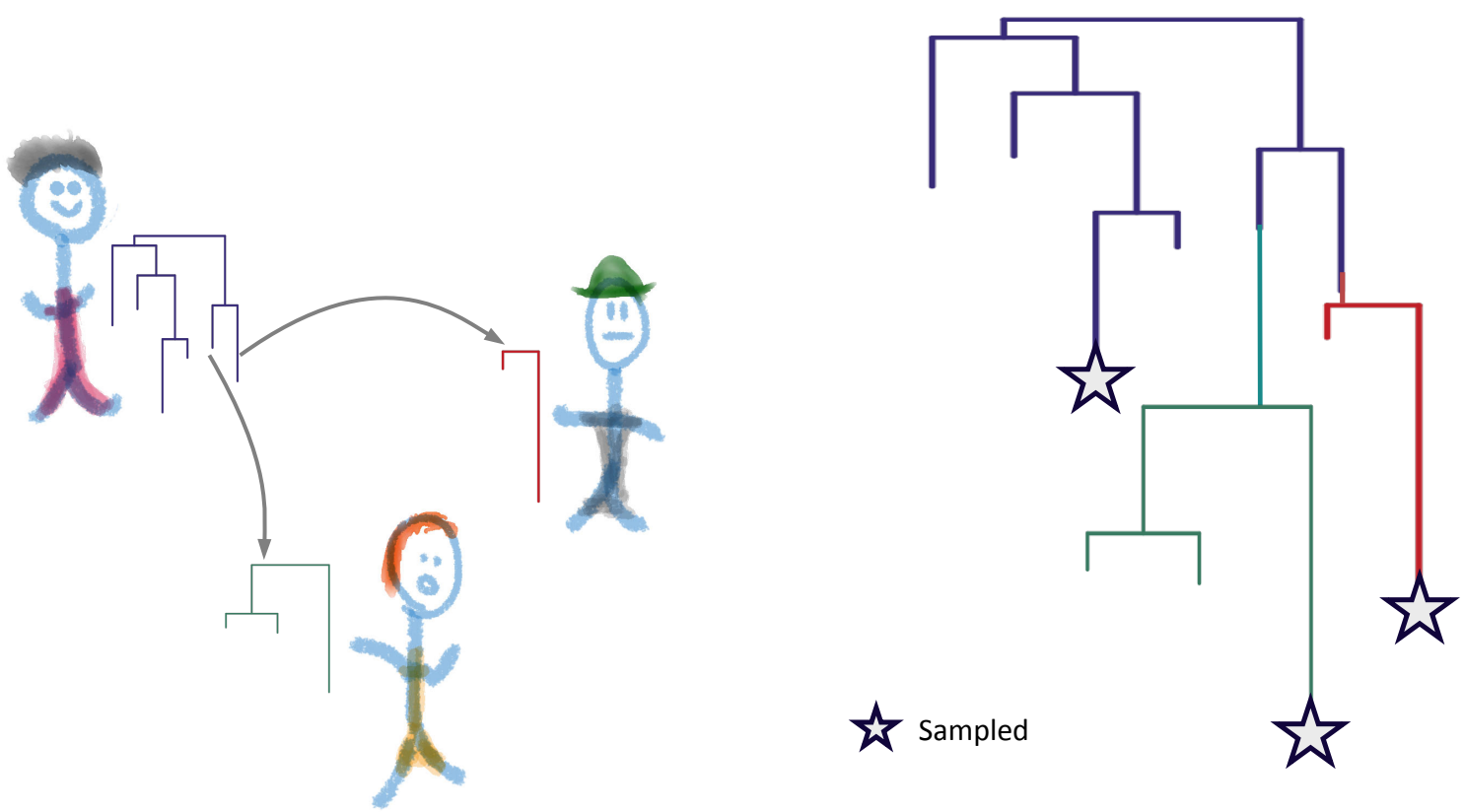


# Genomics doesn't directly reveal transmission

- How closely related are transmission pairs? (variable)
- What if people harbour diverse infections?
- How informative are the times of diagnosis?
- What about unsampled cases?
- What about transmission bottlenecks - big or small?



# In-host diversity



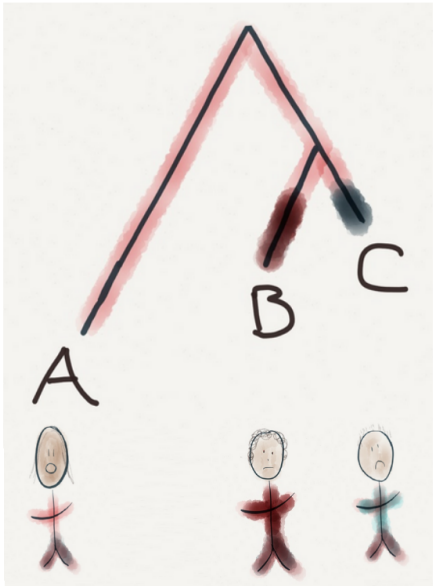

Transmission pairs are not necessarily phylogenetic pairs.

# Colouring can help relate phylogeny and transmission trees

*Lineage*: section of a branch of a tree.

Reasonable constraints:

- Hosts can have more than one lineage at a time
- Each lineage can only be in *one* host at each time
- Lineages change hosts at transmission events.

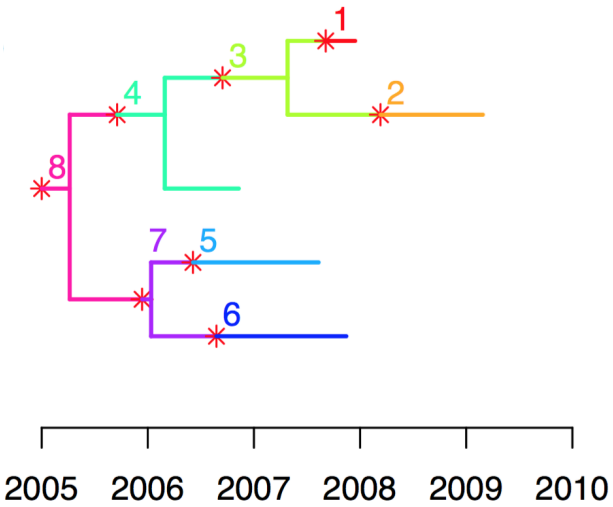



Colour: which host a lineage is in

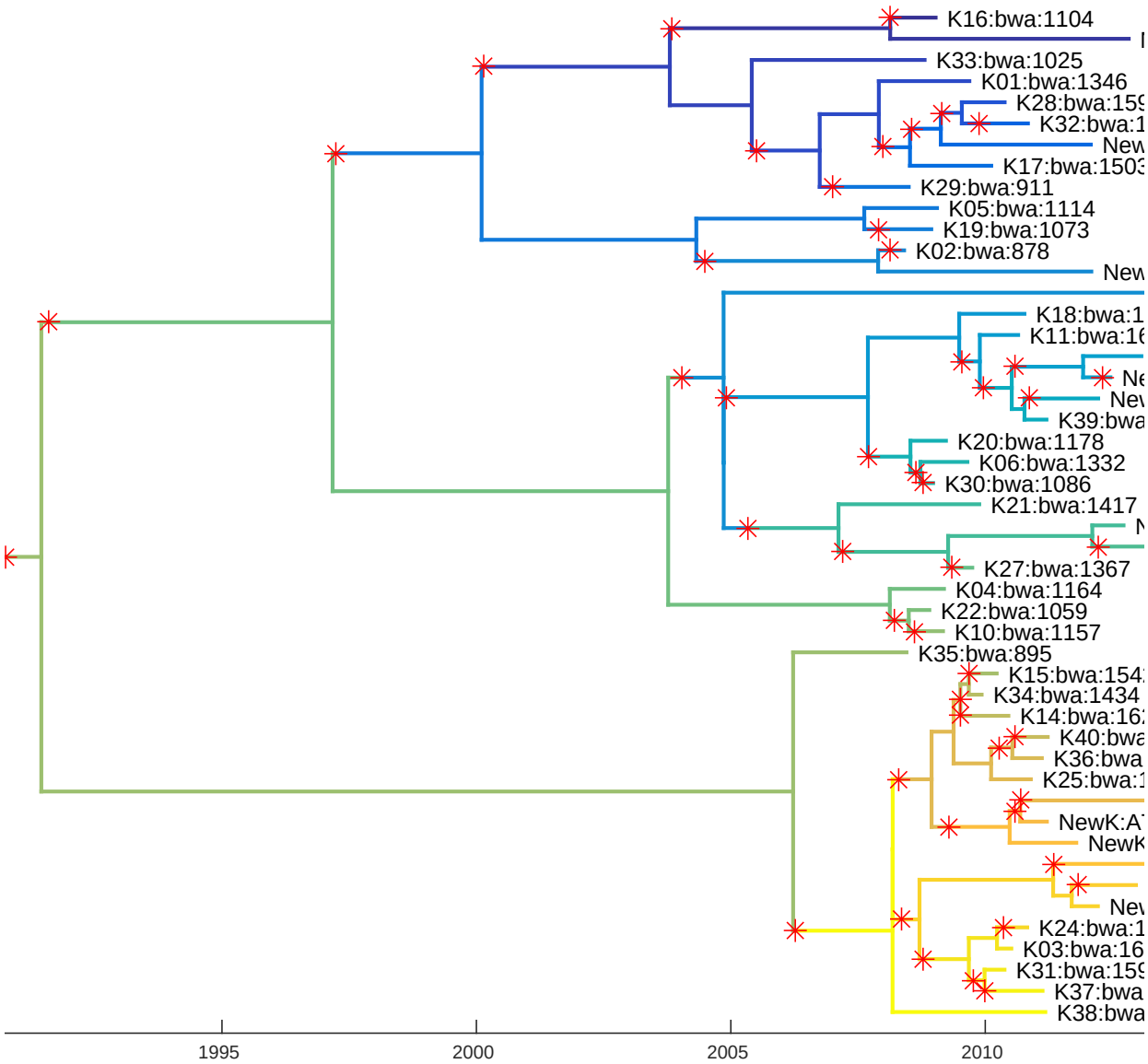
Each admissable colouring corresponds to a transmission tree.

# What is an admissible colouring?

- Each host has a colour
- Not all hosts have to be sampled
- Each lineage is in one host at each time (one colour)
- Colours can't be broken up (each colour must be continuous on the tree)



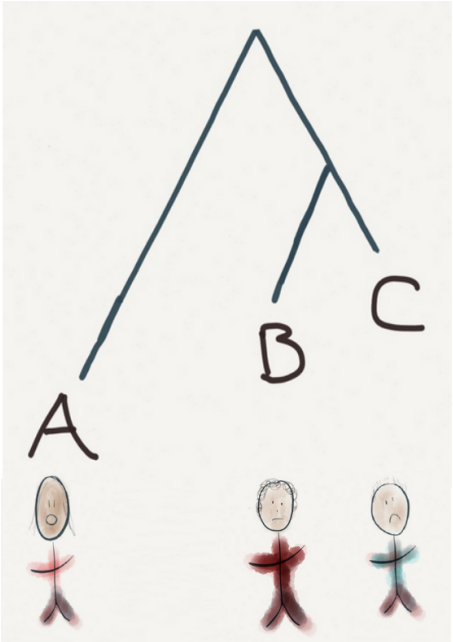
# A valid colouring for a TB outbreak



# How does the phylogeny constrain transmission?

There are constraints!

- If B infected A, B must have infected C
- If A infects B early, then B infected C
- If C infected A, then C infected B




Phylogeny

# We use the colouring to find a likelihood

## Transmission

- *Epi*: epidemiological parameters defining the transmission process
- $T$: transmission tree
- Colour changes are transmission events
- Likelihood: branching process model

## Phylogenies

- $G$: the phylogeny (fixed input from data)
- Transmissions break $G$ into independent $g_i$, one for each host
- We use a coalescent model for $g_i$; coalescent effective population size is $N_e g$

# Decomposition given fixed phylogenetic tree

$$\mathbf{L}(\text{Trans}|\text{Phylo}) \propto \mathbf{L}(\text{Trans events})\mathbf{L}(\text{Phylo}|\text{Trans events})\textit{Priors}$$

**L**(Transmissions):

- Epidemic model for the system: latency, time to infection, time to sampling
- Finite time due to study end (or the present): this modifies the distribution secondary cases depending on infection time and the sampling probability

**L**(Phylo|Trans events) :

- Each colour is independent: many little trees
- Coalescent for each one

# Modelling for the transmission likelihood

- The offspring distribution is negative binomial $(r, p)$. The probability of $k$ offspring is $p_k = \binom{k+r-1}{r-1} p^k (1-p)^r$
- The probability of sampling someone infected at time $t$ is $\pi_t = \pi \int_t^T f_s(\tau - t) d\tau = \int_0^{T-t} f_s(\tau) d\tau$.
- The study ends at time $T$; after that no one is sampled.
- The generation time density is $f_g(\tau)$ where $f_g(0) = 0$ and $\tau$ is the time since infection

# Not all cases are sampled

Let $p_0(t)$ be the probability of being unsampled and having all descendants unsampled, having been infected at time $t$.

As $t \to -\infty$, $p_0(t) \to p_0^*$ which is the solution to the usual equation (conditioning on the number of offspring of the ancestor)

$$p_0^* = (1-\pi)\sum_{k=0}^{\infty} p_k p_0^{*k}$$

However, we have a finite time $T$. We know that $p_0(T) = 1$.

# Math: finite time, unsampled cases

Starting point: probability unsampled and no sampled descendants, if infected at $t$

$$p_0(t) = \text{P(unsampled)} \sum_k \text{P(k offspring at } \tau_j\text{) P(they are unsampled)}$$

$$= (1-\pi_t) \sum_{k=0}^{\infty} p_k(t) \prod_{j=1}^{k} \left[ \int_t^{\infty} f_g^T(\tau_j - t) p_0(\tau_j) d\tau_j \right]$$

$$= (1-\pi_t) \sum_{k=0}^{\infty} p_k(t) [\text{Thing}]^k$$

Now use $p_k$'s generating function, $f(s) = \sum_{k=0}^{\infty} p_k s^k$.

If $p_k$ is negative binomial: $p_0(t) = (1-\pi_t) \left( \frac{1-p}{1-p\text{Thing}} \right)^r$.

# Math: building up transmission likelihood

We solve $p_0(t) = (1 - \pi_t)\left(\frac{1-p}{1-p\ \text{Thing}}\right)^r$ with the trapezoid method, because Thing has $p_0$ in it.

This gives P(unsampled, no sampled descendants | infected at $t$).

Someone in the tree could have infected $k$ others with only $d_0$ of them who are sampled.

$$p(d_0, t) = \sum_{k=d_0}^{\infty} \binom{k}{d_0} p_k \bar{p}_0(t)^{k-d_0} p_s(d_0)$$

which we can compute (typically $d_0$ is small).

# Transmission likelihood

Let host $i$ have: $s_i = 0, 1$ if unsampled, sampled. The times $t^i_{\text{inf}}$ and $t^s_i$ are times of infection, sampling. Then:

$$\mathbf{L}(T|Epi) = \prod_{i=1}^{n} (1-\pi)^{1-s_i} (\pi f_s(t^s_i - t^i_{\text{inf}}))^{s_i} p(d^i_0, t^i_{\text{inf}}) \prod_{j=1}^{d^i_0} f_g(t^j_{\text{inf}} - t^i_{\text{inf}})$$

For each case $i$:

- Was $i$ sampled? likelihood depends on end time $T$ and time of infection $t_i$
- If so, use likelihood for time of sampling for case $i$
- probability ($i$ had $d_0$ *sampled* descendants, ie $p(d_0, t)$
- $\prod_{j=1}^{d_0}$ (likelihood for the time that $i$ had the $j$'th descendant)

# We use the likelihood in MCMC inference

Start with a phylogenetic tree (units of time) and info for the epidemiological model.

1. Propose a colouring: who infected whom, and when
2. Compute its likelihood **L**(Trans|Epi) using the epidemiology model
   - This uses data on how long between infection and sampling, natural history, sampling fraction, basic reproductive number
3. Compute the likelihood for the mini-trees inside each host (coalescent model)
4. Accept or reject the proposal
5. Continue (MCMC)

At the end you have a posterior collection of who infected whom and when transmission trees.

# All together: sequences to transmission

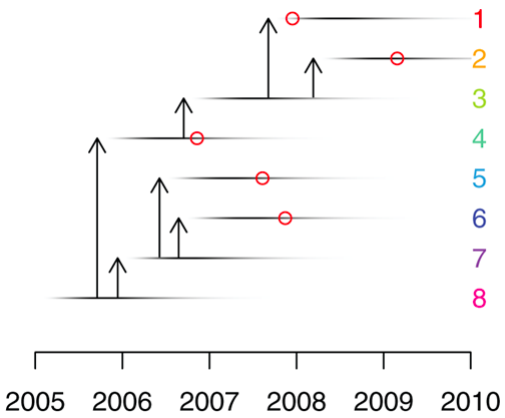
This approach takes in a fixed phylogenetic tree and priors, and produces:

- coloured phylogenetic trees
- transmission trees: who infected whom, and when — useful!
- how long between infection and infecting others — useful!
- how long between infection and sampling — useful!
- placement of missing cases — useful!

Didelot, Fraser, Gardy, Colijn MBE 2017

`TransPhylo`:

https://github.com/xavierdidelot/TransPhylo



# What data does TransPhylo need?

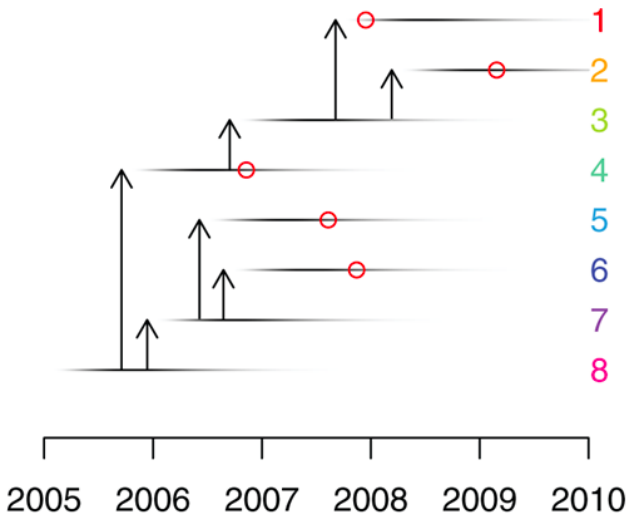
- A timed phylogenetic tree (or a posterior collection of them)
- Sampling dates for the tips (ie the isolates)
- A prior for the time between getting infected and infecting someone else
- A prior for the time between getting infected and getting sampled
- A prior for the overall probability of being sampled eventually
- The time when sampling stopped. Finite time makes a difference! (censoring)

# What does TransPhylo produce?

Formally, TransPhylo estimates 3 key parameters: the mean of the offspring distribution ($R_0$, in epidemiology), the in-host effective population size, and the sampling fraction.

In practice, we use the posterior collection of

- who infected whom?
- generation times
- times between infection and sampling
- unsampled cases and their locations in the phylogeny



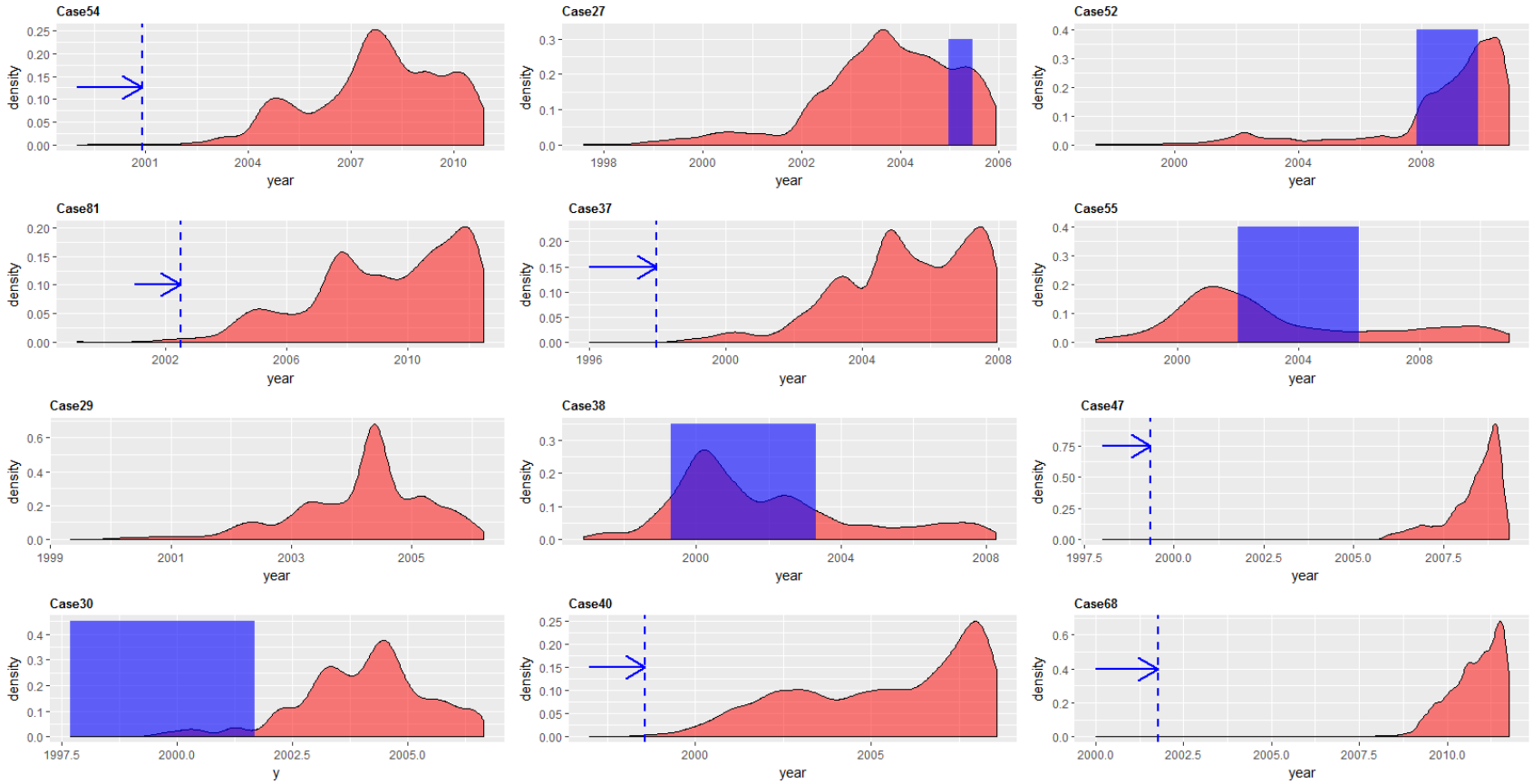
# TB clusters in Norway

- We analyzed a strain of TB circulating in Norway
- Often, a case's country of origin has higher TB burden than Norway
- Closely-related cases suggested recent transmission.
- Generation time and sampling time priors reflect uncertainty
- Question: is there transmission in Norway?

*Genome-based transmission modelling separates imported tuberculosis from recent transmission within an immigrant population.*

Ayabina et al, Microbial Genomics, 2018

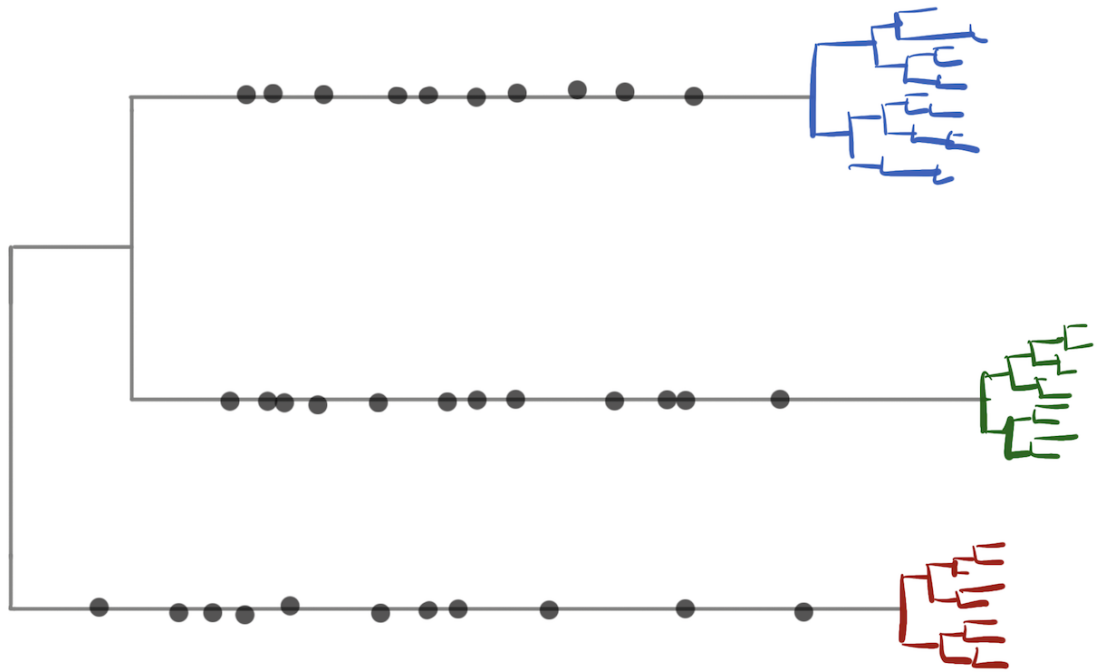
# Infected in Norway or not?




Red: Posterior time of infection. Blue: arrival in Norway.

Some cases were very likely infected in Norway.

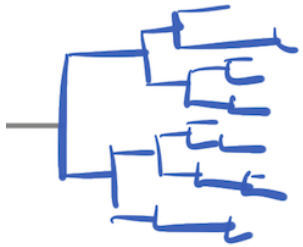
# What if the data are not just one big tree?

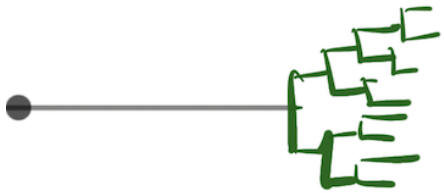


- Many small clusters, large genetic distances away
- TransPhylo has to place unsampled cases on the long branches
- it will not explore transmissions on clusters efficiently.

# Many input trees

- There may be many input phylogenetic trees
- boostrapping, different clock rates, tree models
- Bayesian tree inference

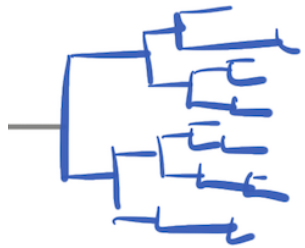


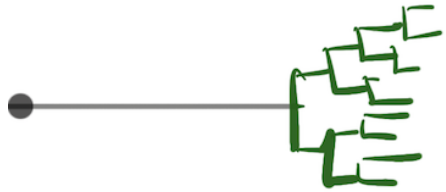


**Improvement**: use TransPhylo on *lots of trees* at the same time, *sharing parameters* between them.

# Idea: share parameters from one dataset to another

- Simultaneous transmission inference
- TransPhylo formally estimates 3 things: $R_0$, sampling fraction, in-host effective population size
- For $n$ clusters separately, that would be $3n$ parameters
- With parameter sharing: estimate 3 parameters instead of $3n$.
- Each cluster is informed by the others.
- Computational efficiency, tighter CIs





# Joint transmission inference

- Joint tree space: Let $T = (T_1, \ldots, T_n)$ where $T_i$ are the individual clusters (trees)
- Likelihood
  - no parameter sharing - need to estimate all the $\theta_i$:

    $p(T|\theta) = \prod_{i=1}^{n} p(T_i|\theta_1, \theta_2, \theta_3, \ldots \theta_n)$,
  - with parameter sharing - need to estimate only $\theta$:

    $p(T|\theta) = \prod_{i=1}^{n} p(T_i|\theta)$

  where $\theta$ is one set of parameters common for all clusters.
  We can choose to share some or all of the parameters.

# Case study of TB transmission clusters

- 110 transmission clusters of TB in Valencia, Spain from 2014.
- We have sequence alignments and sampling dates; 764 cases.

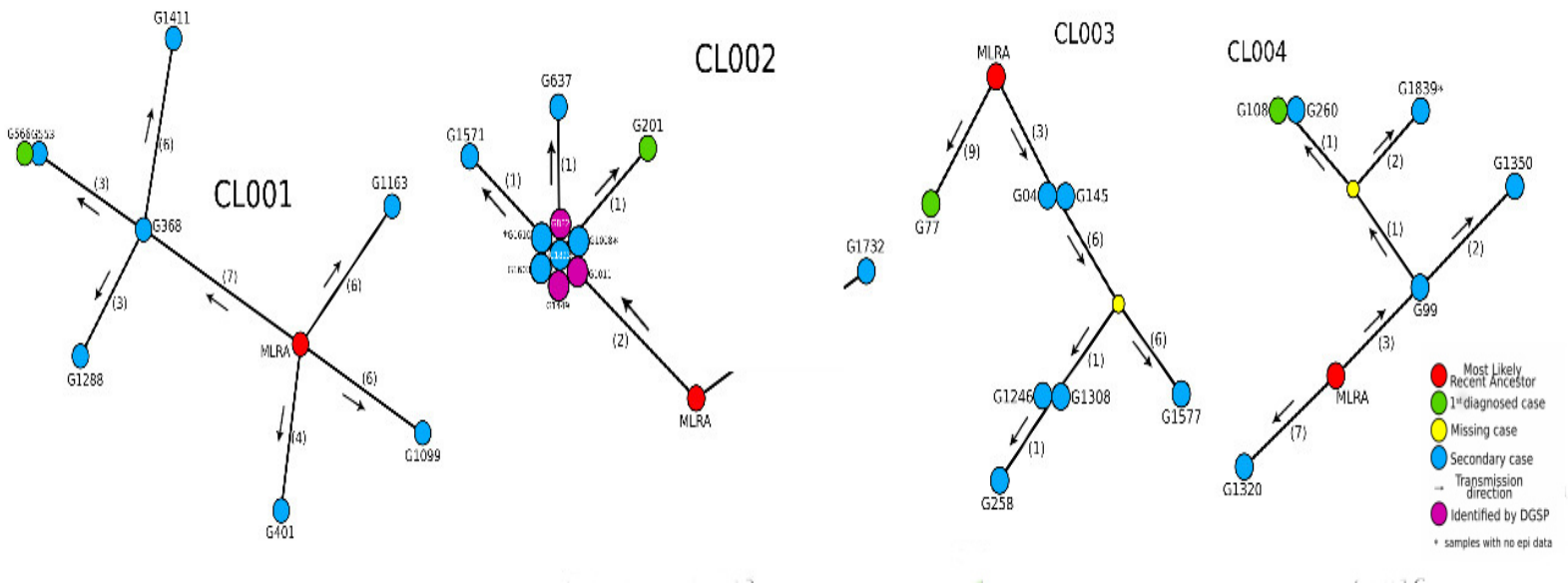


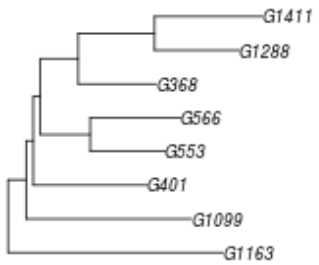

Figure: Genetically-defined transmission clusters (from Iñaki Comas, TB Genomics Unit, Valencia). Xu et al, *High-resolution mapping of TB transmission...* , PLOS Medicine, 2019.

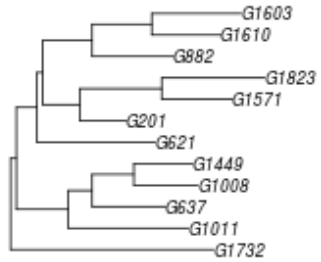
# Questions from public health

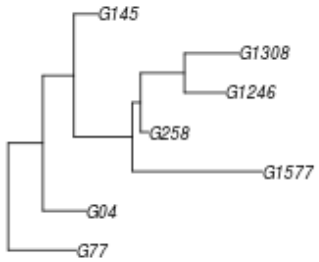
- Was the first diagnosed case the index case?
- Is there evidence of transmission before (recalled, diagnosed) symptom onset?
- How long between infection and infecting others?
- How long between infection and sampling?
- Are most clusters chains, with each case infecting one other? or more complex patterns?
- How many cases are we missing? In particular, are we missing the index case?

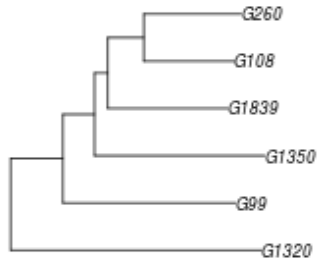
# The dated phylogeny for clusters CL001-CL004

- Use `treedater` to date the phylogenetic trees
- Treedater can use time intervals for tips without exact dates

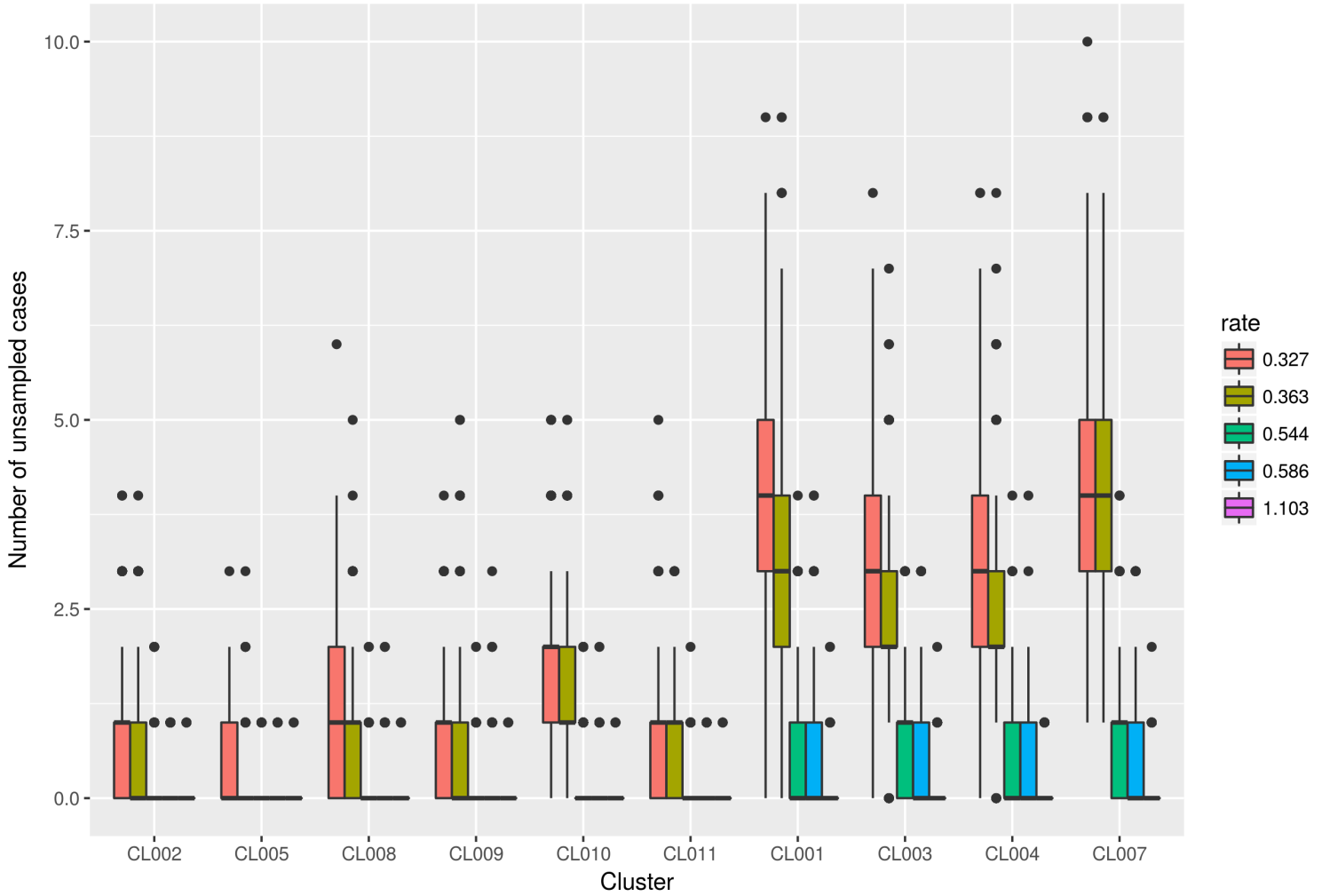


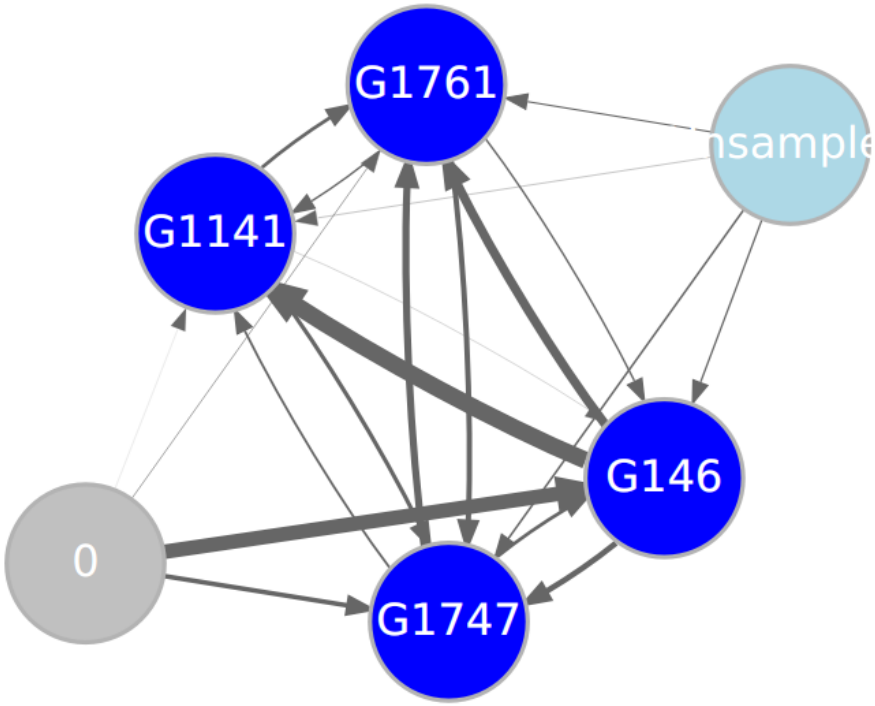



# Some clusters have more unsampled cases than others



# Infection events are uncertain

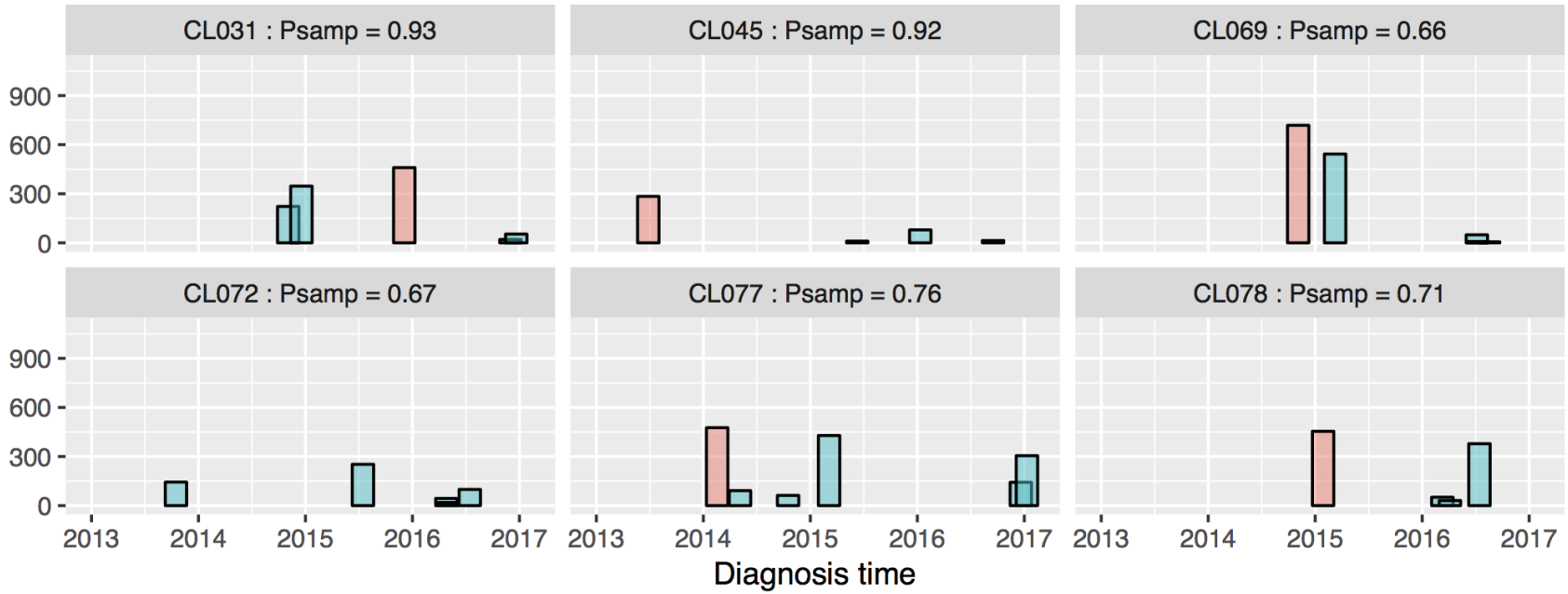
Infection network for cluster 45 showing uncertain infection events




Width: posterior probability

We can separate likely TB transmitters from other clustered cases.

# Was the first diagnosed case the index case?



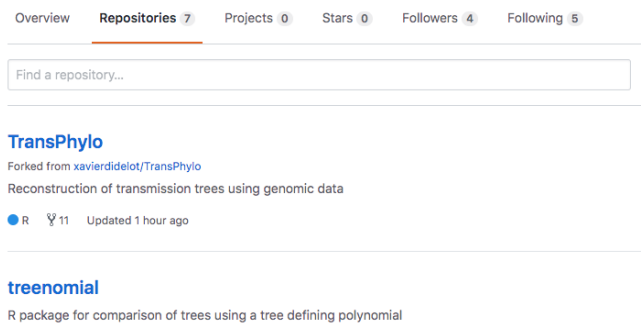

Height: number posterior trees where the case was the index case.
Red: most likely index case

# Incorporating epidemiological data

We can include information in the prior. This helps resolve transmission.

- Location
- Plausible transmission time (eg negative screen, leaving date)
- Symptom time
- Contact data: when two individuals were in contact

Matthew Gould
mattgou1d




`https://github.com/mattgou1d/TransPhylo`

# Advantages of the TransPhylo approach

- Includes genetic data and unsampled cases
- Transmission trees consistent with phylogeny - respect constraints
- Posterior collection of augmented trees - easy to interpret
- Captures uncertainty - considerable uncertainty left even with genomic data
- Can suggest where there are unsampled infectors, infectees
- Can distinguish *credible transmitters* from all clustered cases
- Can scale to large-ish data sets using multiple trees simultaneously

# Limitations: transmission

We model person-to-person transmission, dense sampling

- No environmental reservoir - sinks, taps, kitchens, ponds, vectors etc
- Models direct transmission events. If you have $<\sim 75\%$ sampling, another method is likely to be better
- Currently: fixed sampling over time (quite easy to change)

It is a two-stage process (first phylogeny, then transmission):

- Requires timed phylogenetic tree(s)
- Constructing these can be cumbersome and noisy
- Simultaneous reconstruction of the phylogenetic tree *and* the transmission tree? (hard but do-able)
- If sequences don't define one true phylogenetic tree, using many combined is OK but simultaneous inference likely better

# Limitations: covariate data

- The branching process likelihood uses a recursion where we condition on the $k$ infectees, of whom $k - d_0$ were unsampled.
- We integrated out the unsampled cases' unknown times of infection
- We can't integrate out their unknown values of lots of covariates: location, demographics, clinical history, prison history etc etc
- Best we can do: eg penalize transmission trees that "break rules", by setting priors for individual transmission events based on covariate data
- Perhaps a survival analysis likelihood would do better: see Eben Kenah's talk, and work

# Thank you. Questions?

- Matthew Gould (SFU), Jessica Stockdale (SFU)
- Vegard Eldholm (Norway)
- Yuanwei Xu (Imperial)
- Iñaki Comas (Valencia)
- Xavier Didelot (Warwick), Christophe Fraser (Oxford), Jennifer Gardy (Gates Foundation)