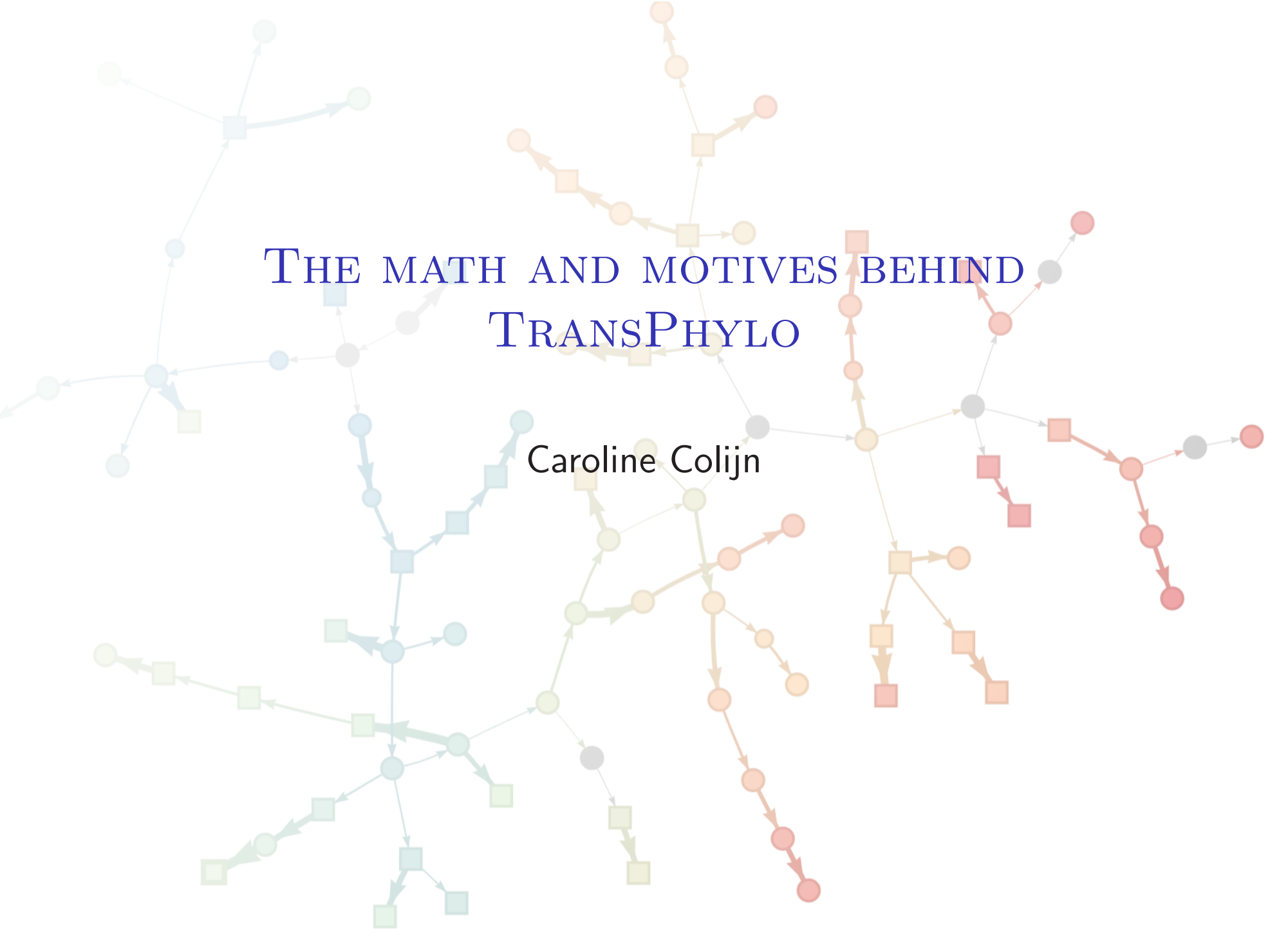
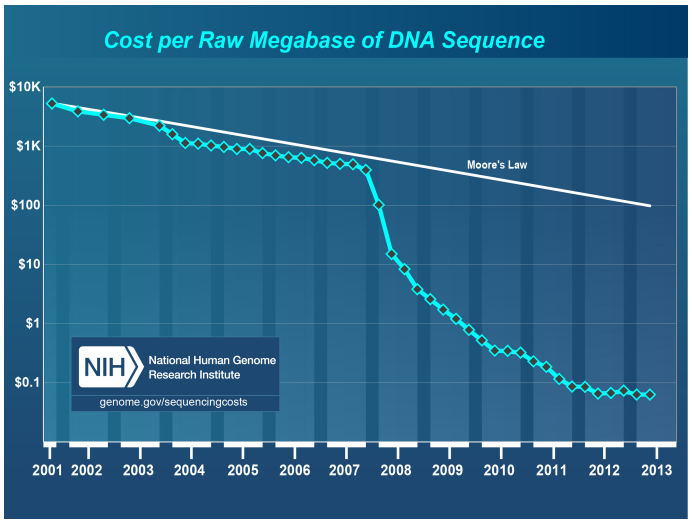


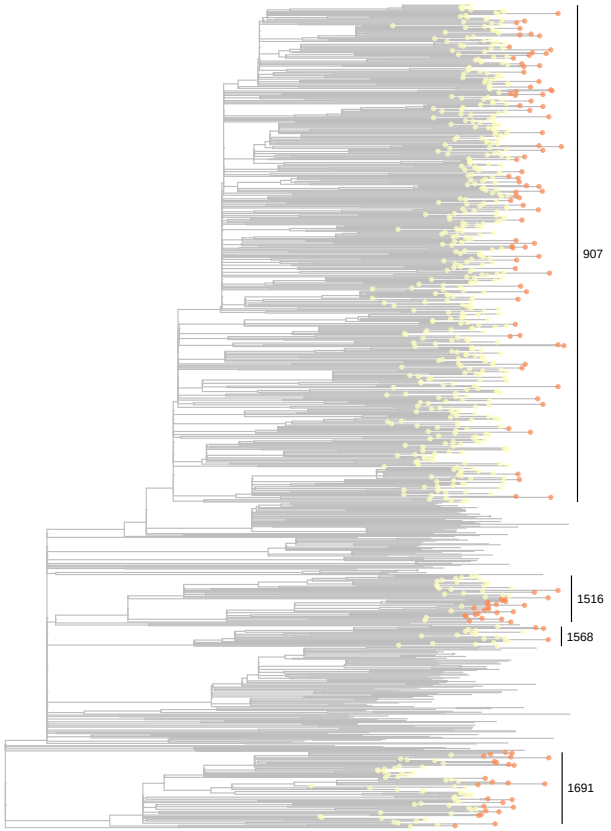
The math and motives behind TransPhylo

Caroline Colijn

Sequencing for pathogens: there is a lot of data



Sequencing is cheap now



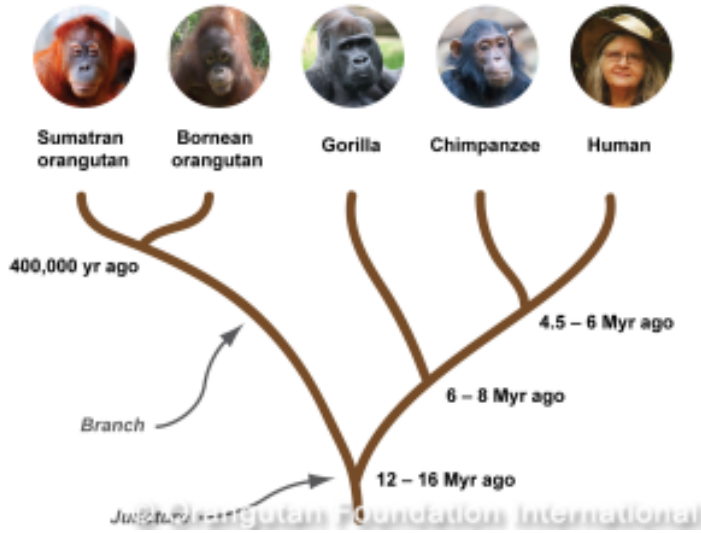
Recent influenza H3N2 subset

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

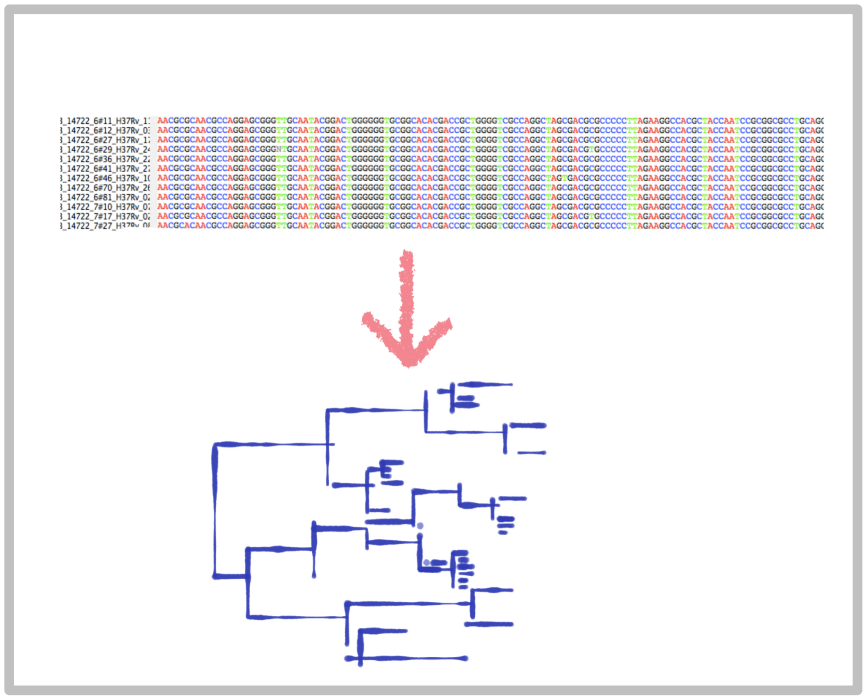
Trees describe patterns of ancestry through time

We understand sequence data using trees.

Definition: A *phylogeny* is a (rooted, binary) tree in which tips correspond to organisms and other nodes correspond to common ancestors.



We have sequences

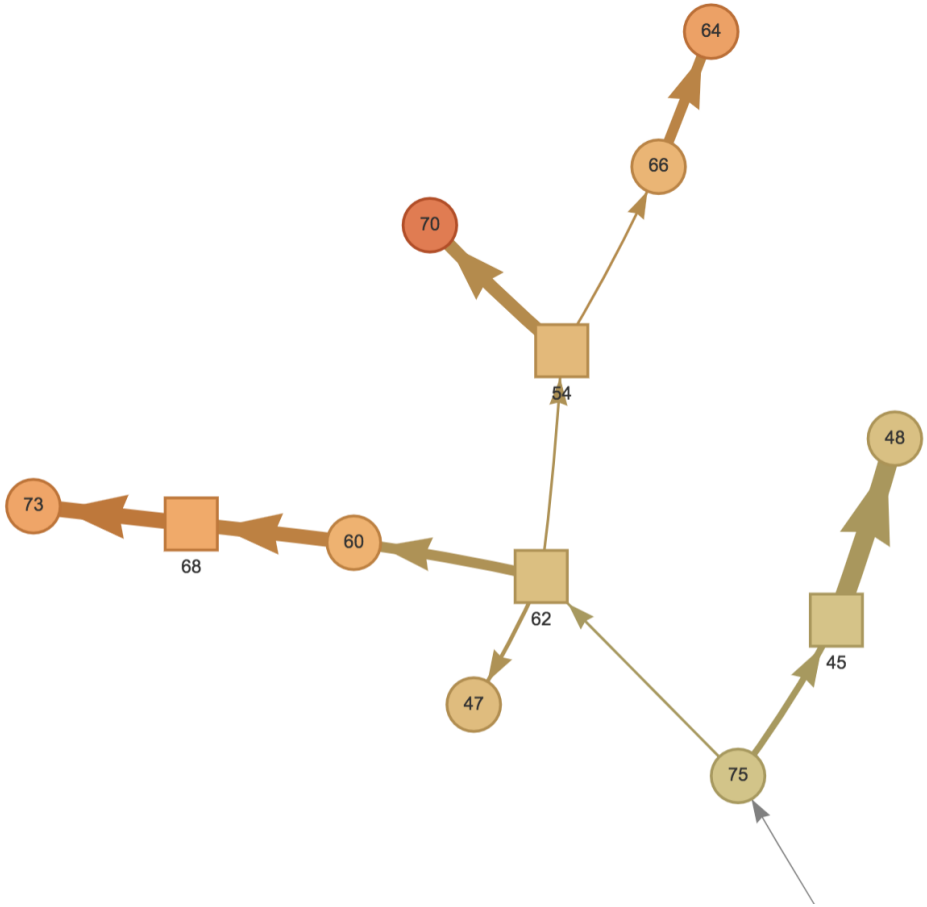


REMEMBER THE OLD GAME OF "TELEPHONE?

Transmission tree

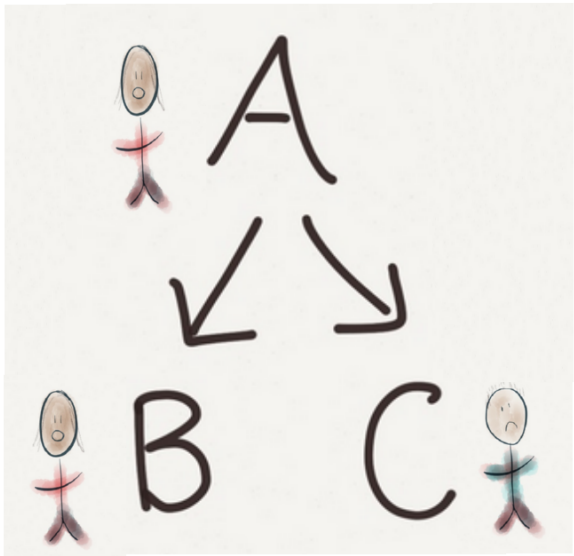
Definition: A *transmission tree* is a tree in which nodes are people and edges (directed) correspond to infection events.

Edges may be associated with times of infection.

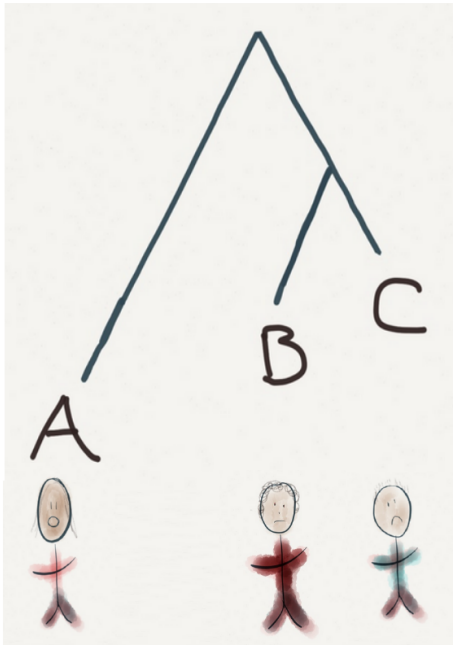


TransPhylo: Use phylogeny to understand transmission

Example: transmission tree and phylogeny



A infects B and C



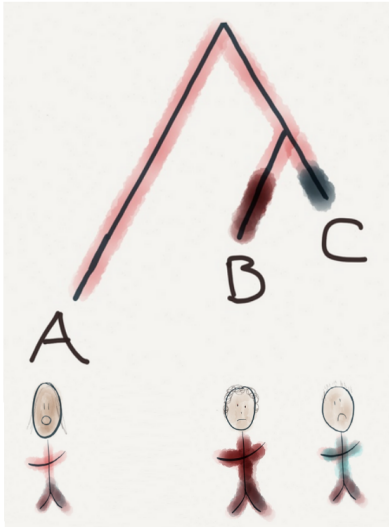
Phylogeny

Colour the phylogeny

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.

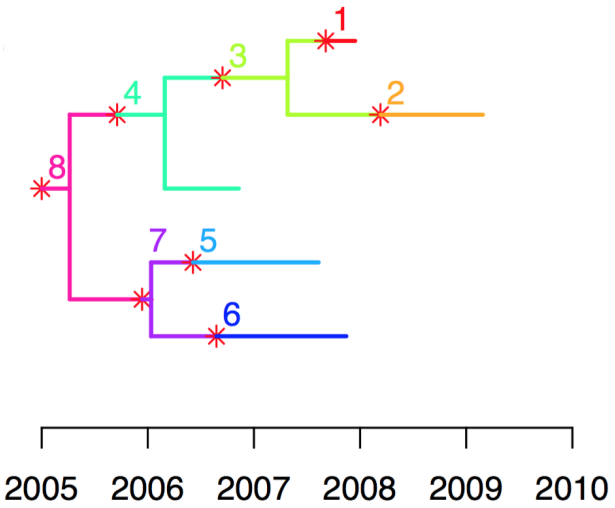


Colour: which host a lineage is in

Each admissable colouring corresponds to a transmission tree.

What is an admissible colouring?

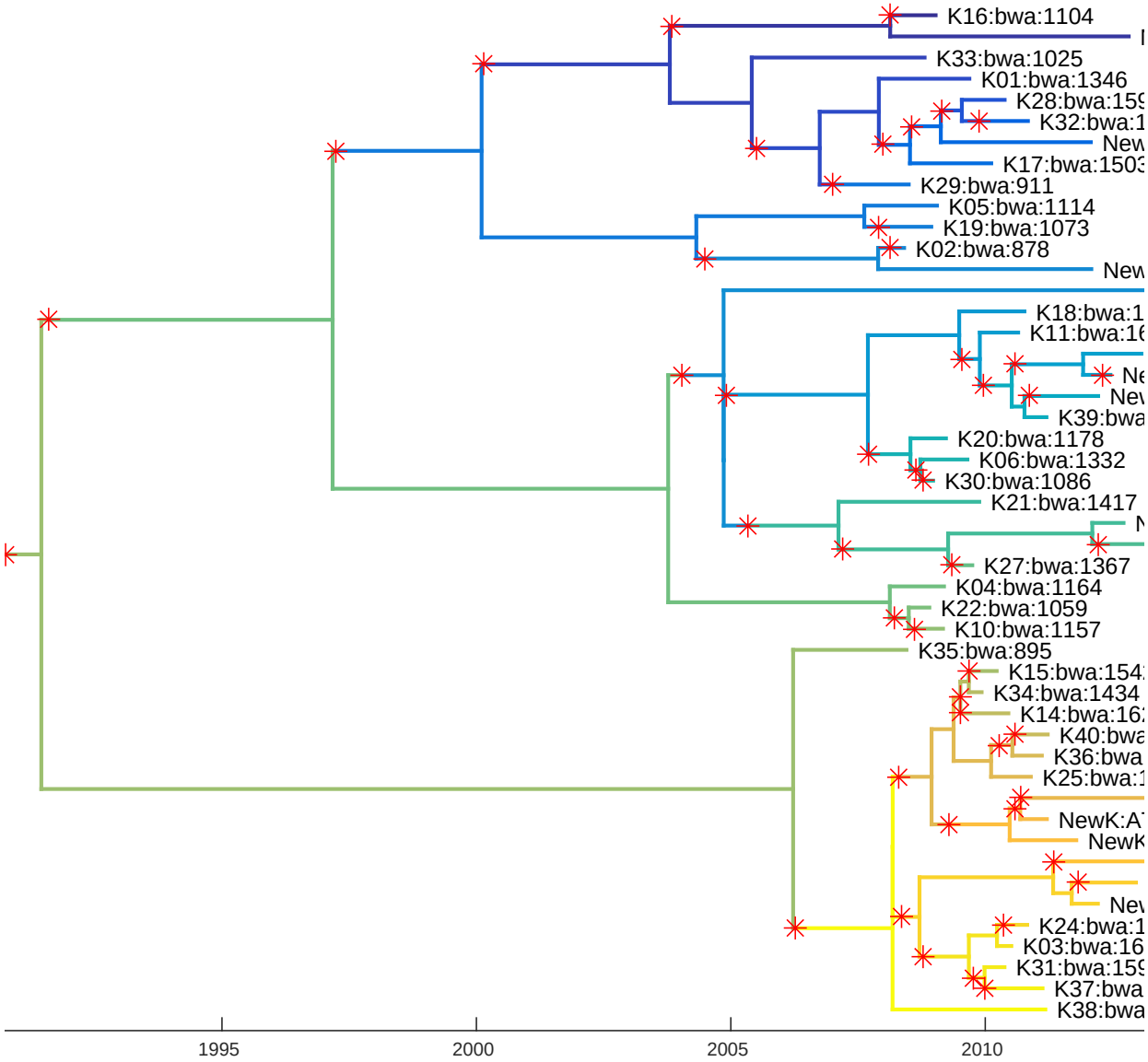
- Each host has a colour
- Not all hosts have to be sampled
- Each lineage is in one host at each time (one colour)
- Colours can't be broken up (each colour must be continuous on the tree)



Admissible colouring:

Each tip has its own colour; colour doesn't extend after the tip recovers; colours are connected

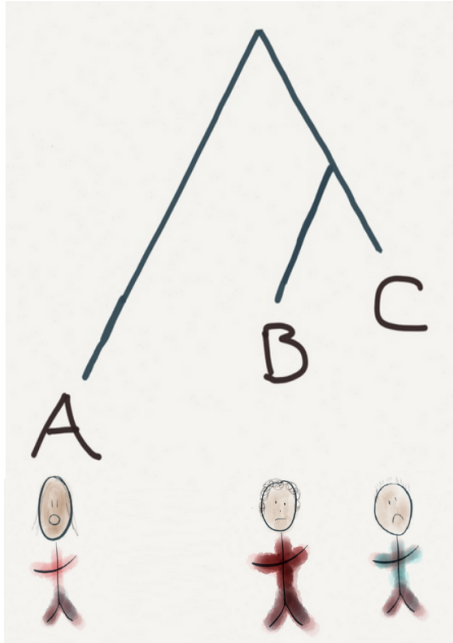
A valid colouring for a TB outbreak in Kelowna, BC



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

Bayesian approach

TransPhylo is a two stage approach:

1. Make a timed phylogenetic tree (later, use many of them), G
2. Layer transmission events (transmission tree T) on top of it
3. θ denotes the parameters of the epidemic model
4. $N_e g$: coalescent parameter

Bayes' theorem gives us (with L = likelihood)

$$\mathrm{L}(Epi, N_e g, T|G) \propto \mathrm{L}(G|Epi, N_e g, T)\mathrm{L}(Epi, N_e g, T)$$

$$= \mathrm{L}(G|N_e g, T)\mathrm{L}(T|Epi)\mathrm{L}(Epi)\mathrm{L}(N_e g)$$

In words: the transmission tree chops up the phylogeny into little bits, giving $\mathrm{L}(G|N_e g, T)$. The transmission's likelihood depends on the epidemiology.

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

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

L(Transmissions):

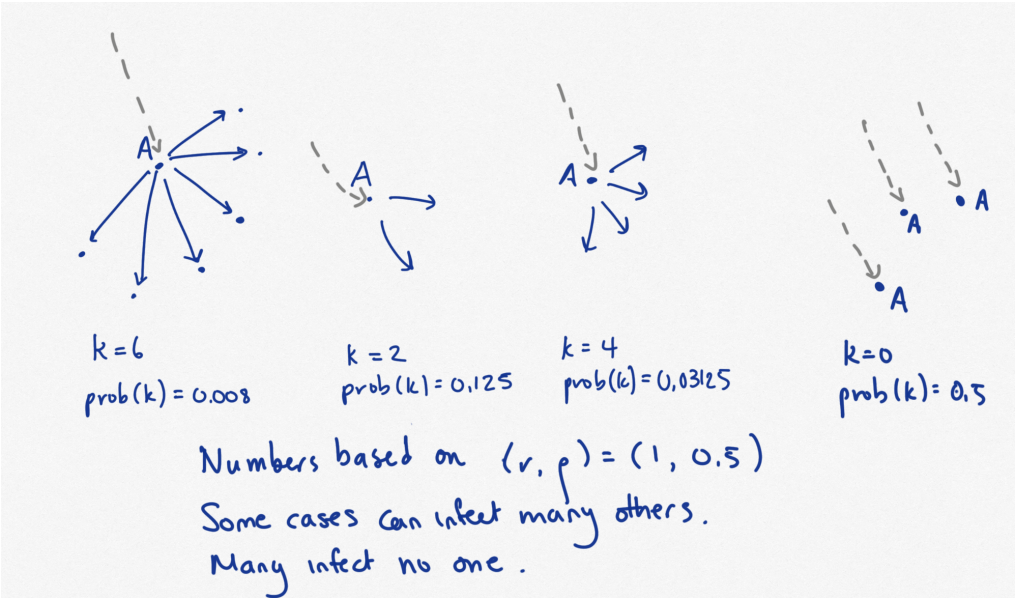
- Epidemic model for the system: latency, time to infection, time to sampling
- Finite time due to study end (or the present): this modifies the distribution secondary cases depending on infection time and the sampling probability

L(Phylo|Trans events) :

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

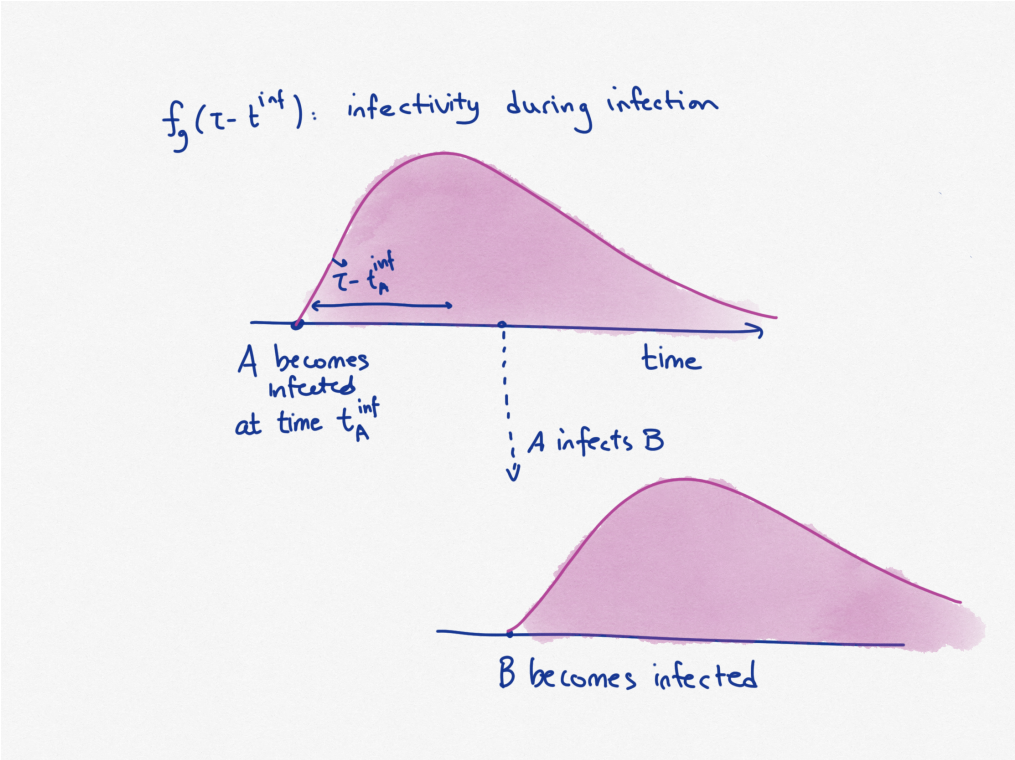
The epidemiological model ("Epi") - 1: how many secondary infections?

- We use a negative binomial (r, ρ) distribution for the number of secondary cases
- The probability of k offspring is $p(k) = \binom{k+r-1}{r-1}\rho^k(1-\rho)^r$



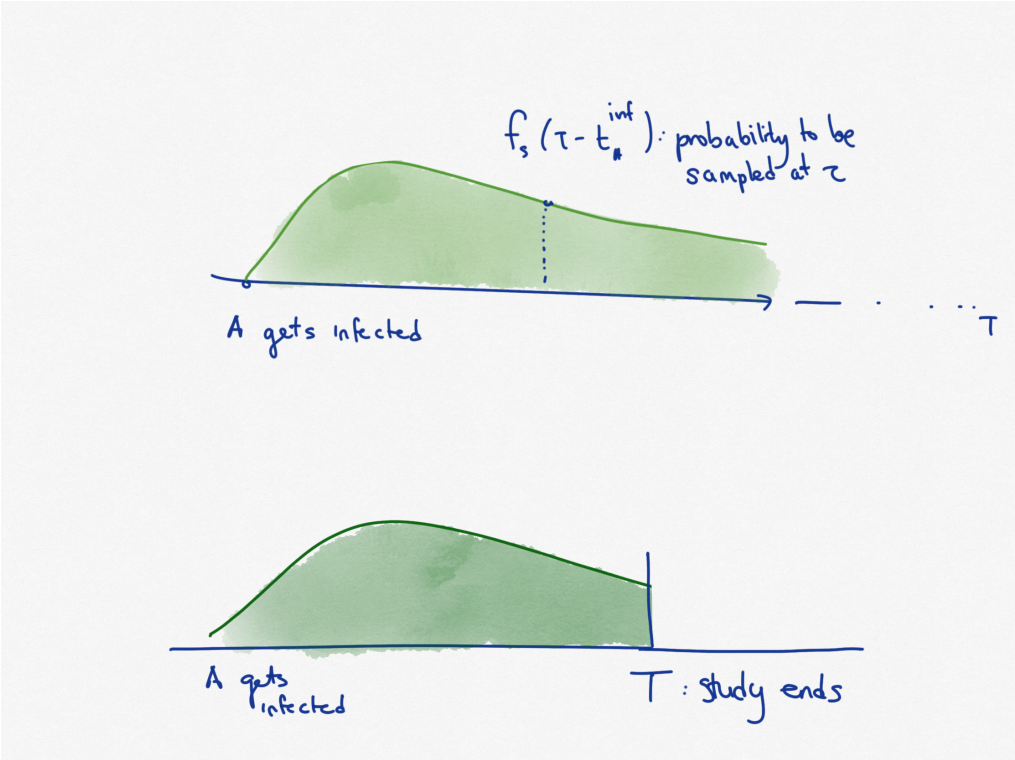
Time to infect others

- The generation time density is $f_g(\tau)$ where $f_g(0) = 0$ and τ is the time since infection



Sampling

- The probability of sampling someone infected at time t is $p_s(t) = p_s \int_t^T f_s(\tau - t)d\tau = \int_0^{T-t} f_s(\tau)d\tau$.
- The study ends at time T; after that no one is sampled.



Probabilities for unknown unknowns

Let $p_0(t)$ be the probability of being unsampled and having all descendants unsampled, having been infected at time t. Suppose the outbreak started a very long time ago ($t \to -\infty$), and $p_0(-\infty) = p_o^*$. Then p_0^* should solve this equation :

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

In words: for you to be unsampled and have no sampled descendants:

1. you have to be unsampled: probability $(1 - p_s)$
2. maybe you had k descendants. They all have to be unsampled (with no sampled descendants) too

What about when the outbreak was NOT a very long time ago?

In order to be unsampled and have no sampled descendants:

- You need to be unsampled: probability $1 - p_s(t)$
- All of your k descendants also need to be unsampled with no sampled descendants.
- But we don't know when you infected them! Now it matters, because if you infected them yesterday and the study ends today, they will not have been sampled. This impacts $p_0(t)$.
- We integrate out the uncertain time of infection of the secondary cases. Say the infection was at time τ_j
- The probability of this is $f_g(\tau_j - t)$, AND this new infectee has to be unsampled with no sampled descendants
- So there is is a term for each descendant: $\int_t^\infty f_g(\tau_j - t)p_0(\tau_j)d\tau_j$ instead of p_0^*

Probability of no descendants: finite time

Integrating out unknown times, we have

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

Let the term in square brackets be $\bar{p}_0(t)$. There are k of these, and they are all the same. We have

$$p_0(t) = (1 - p_s(t)) \sum_{k=0}^{\infty} p(k) \left[\int_t^{\infty} f_g(\tau_j - t) p_0(\tau_j) d\tau_j \right]^k$$

$$= (1 - p_s(t)) \sum_{k=0}^{\infty} p(k) \bar{p}_0(t)^k \quad (2)$$

Use generating function

- Generating functions are a very convenient way to handle sums like this.
- Definition: $g(s) = \sum_{k=0}^{\infty} p(k)s^k$.
- We know a LOT about generating functions, including the form of $g(s)$ for common distributions like the negative binomial
- Negative binomial: $g(s) = \left(\frac{1-\rho}{1-\rho s}\right)^r$

The previous slide's equation becomes the *integral equation*

$$p_0(t) = (1 - p_s(t)) \left(\frac{1-\rho}{1-\rho \bar{p}_0(t)}\right)^r .$$

We solve it with the trapezoid method and the assumption $f_g(0) = 0$. .

Probability $p(d_0)$ sampled descendants

So we know the probability of having no sampled descendants, if infected at time t.

Now we condition on the total number of descendants; choose 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 and higher k terms vanish quickly in k).

Transmission likelihood: components

Now we can build the likelihood for the transmission tree.

For each case i, we use:

- 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)

Transmission likelihood

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

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

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)

Put it all together

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

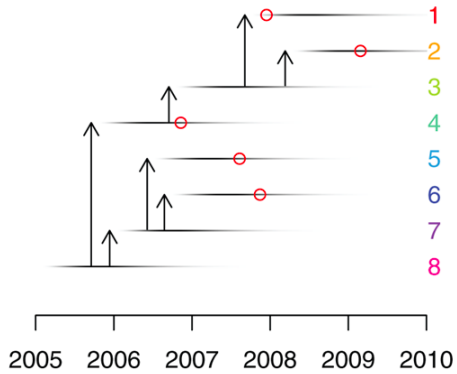
This approach takes in a fixed phylogenetic tree and priors, and produces:

- 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?

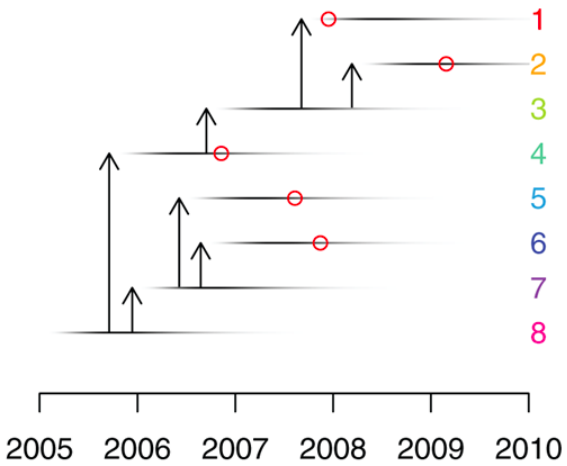
- A timed phylogenetic tree (or a posterior collection of them)
- Sampling dates for the tips (ie the isolates)
- A prior for the time between getting infected and infecting someone else
- A prior for the time between getting infected and getting sampled
- A prior for the overall probability of being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

What does TransPhylo produce?

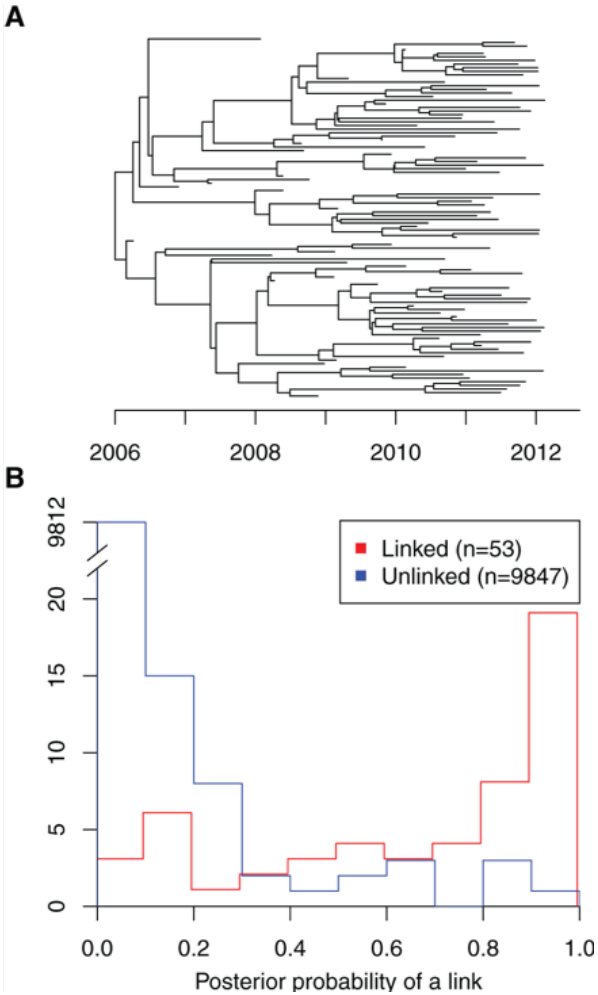
Formally, TransPhylo estimates 3 key parameters: the mean of the offspring distribution (R_0, in epidemiology), the in-host effective population size, and the sampling fraction.

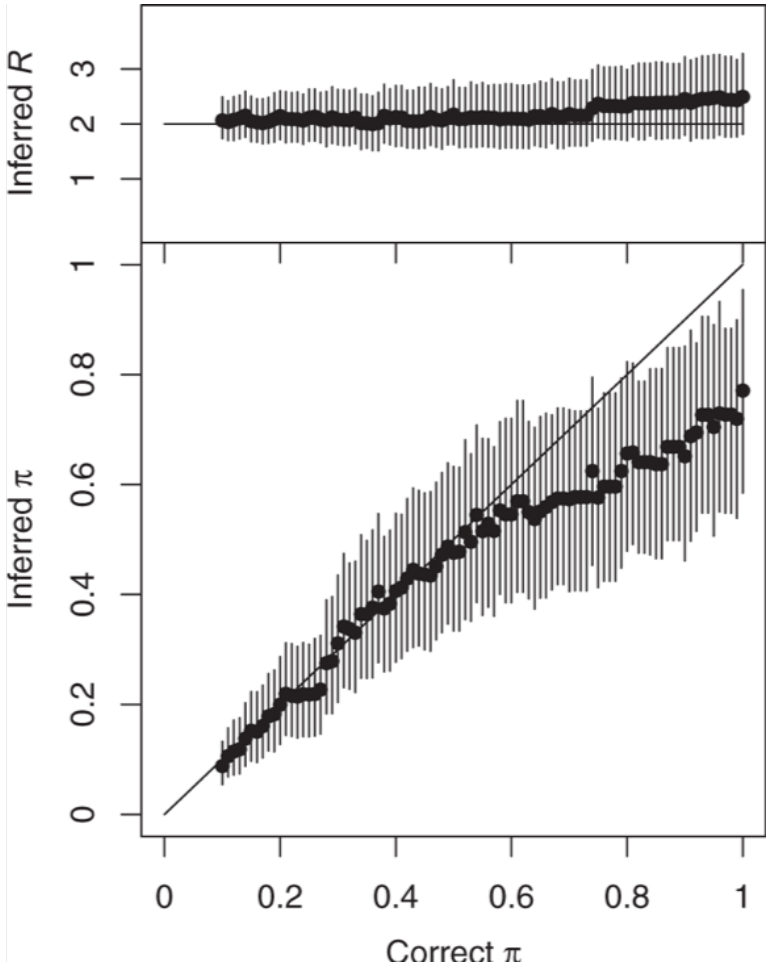
In practice, we use the posterior collection of

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



Performance





Didelot et al, MBE, 2017: Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks

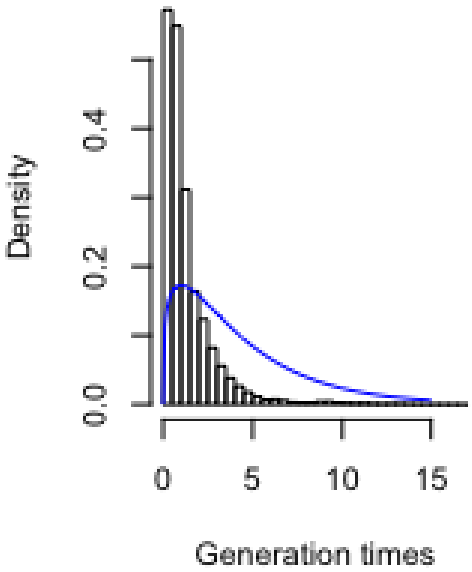
A 13-year TB outbreak in Hamburg

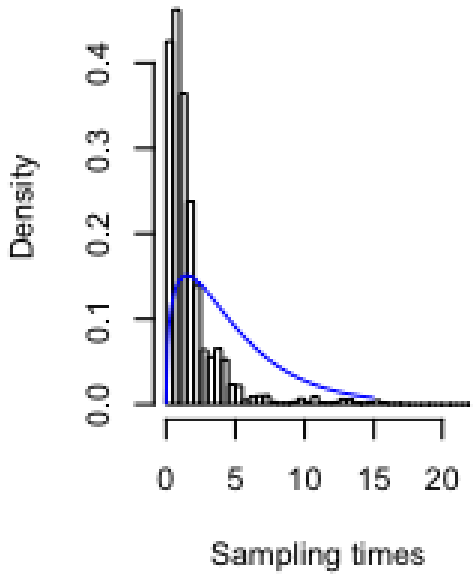
- Outbreak of 86 tuberculosis cases over 13 years. Roetzer et al 2013.
- Active case finding among contacts of cases
- Cases also identified for reasons other than TB infection
- Generation time and sampling time priors reflect uncertainty

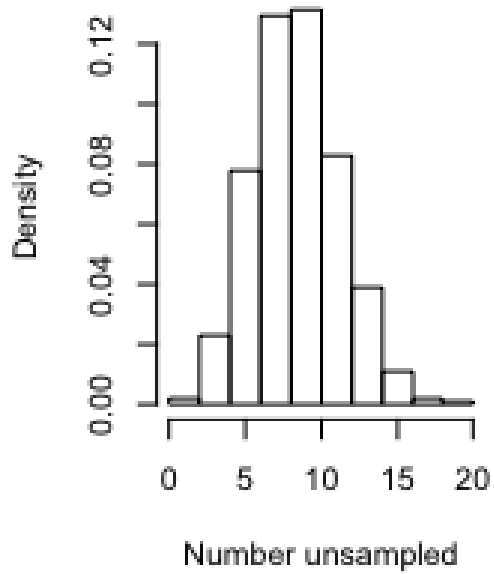
Time to infection, time to sampling, number unsampled

Cases infected someone within 2 years (80%) *among transmitting cases*. 75% sampled in 2 years.

It is likely that some cases were unsampled.

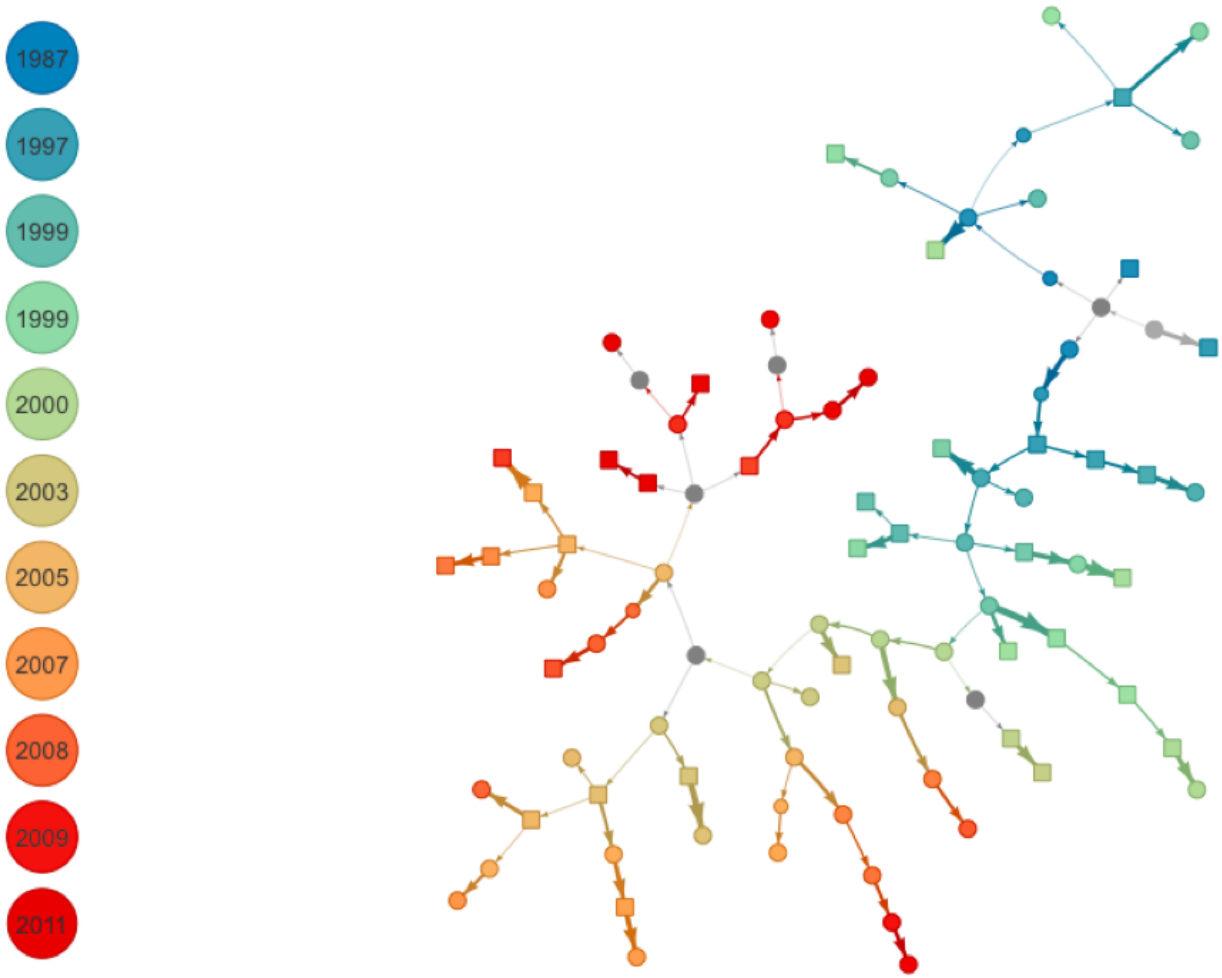






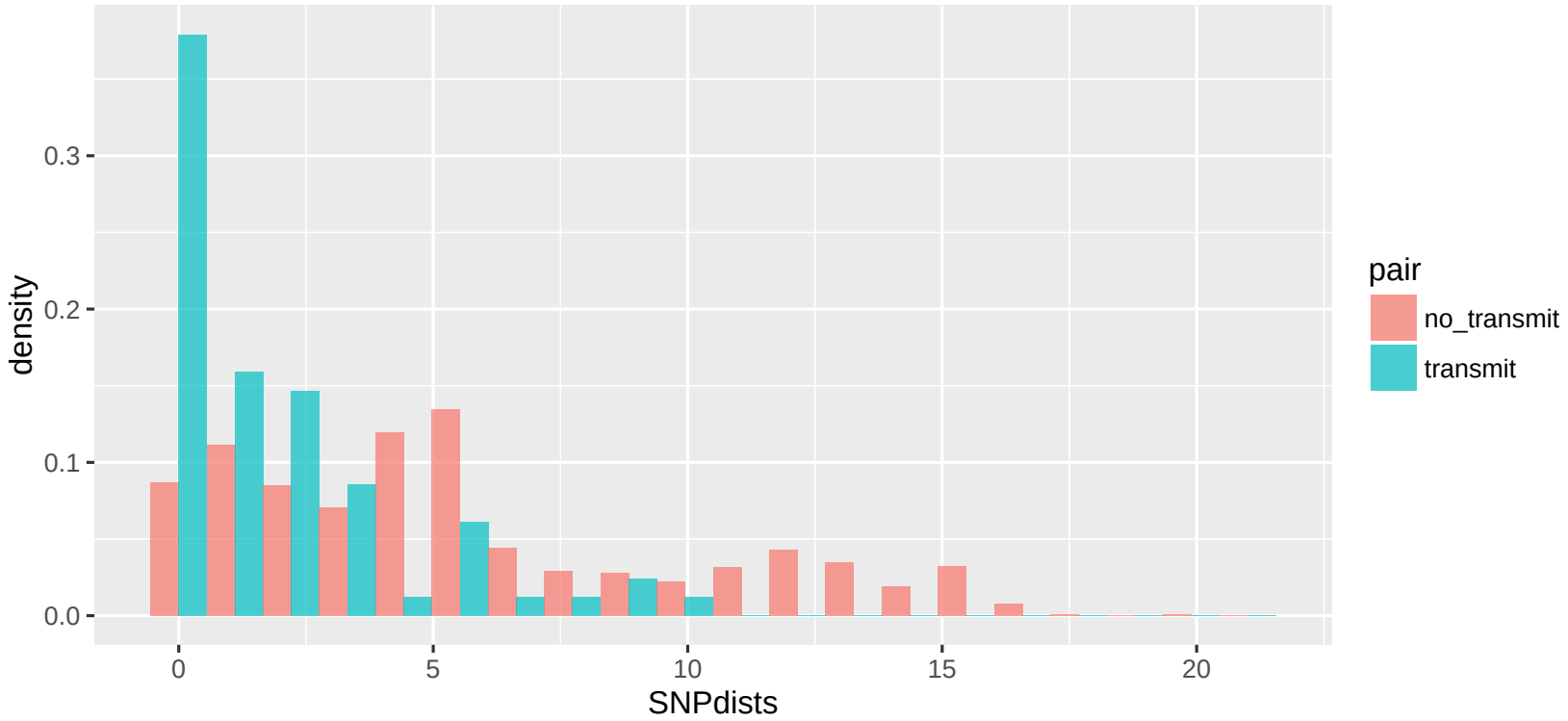
Prior in blue

One transmission tree to summarise them all



width: posterior prob. length: SNP distance. colour: time of infection.
grey: unsampled. square: smear-positive

SNP distances between inferred transmission pairs



When and where was each case infected?

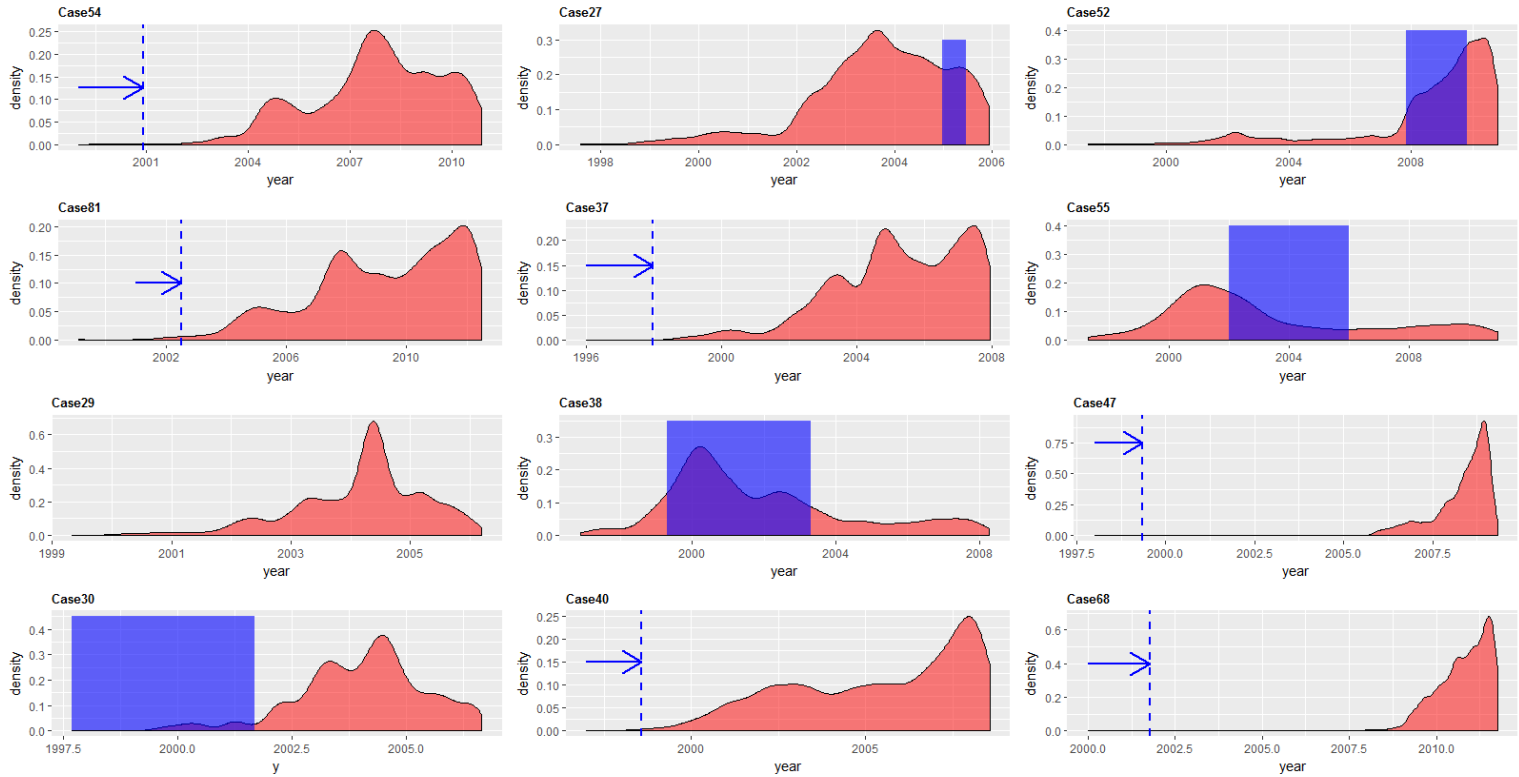
TransPhylo combines the priors for generation and sampling time, plus the genetic data, to give posterior times of infection for each case.

Data:

- Cluster of closely-related cases in Norway
- Cases occur among people immigrating to Norway
- It is often assumed that they were infected before arriving
- But genomic data show signs of recent transmission

We compared time of arrival to posterior time of infection for 13 closely-related cases

Infected in Norway or not?



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

Some cases were very likely infected in Norway.

Ayabina et al, Microb. Genom.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6249437/

Strengths of this approach

- Each transmission tree is consistent with the phylogeny
- In-host diversity is allowed, and accounted for
- Other approaches limit the possible transmission trees:
  - not enough: pairwise methods
  - too much: methods that assume that branching events in the phylogeny are the same as transmission events
- Good treatment of the sampling process
- Flexible epidemiological model

Limitations and extensions

- The two-stage process: timed phylogeny first, then layer transmission on top of it
  - Work by Yuanwei Xu: use many phylogenies, not just one. See Xu et al, PLOS Medicine 2019
  - But is that really better than say outbreaker?
- Challenging to bring in additional data, due to the way we treat unsampled cases
  - Extension: use priors on the transmission tree
  - Extension: connect posterior transmission trees to extra data for patients
  - For example, what could predict whether someone is a "credible infector"?

A few more limitations

- Does not handle multiple infections, reinfection. Each host is assumed to be infected precisely once.
- Assumes that a single pathogen infects each person (bottleneck of 1)
- Does not infer phylogeny and transmission simultaneously
  - BEASTLIER (Matthew Hall) does, similar model, but no unsampled cases
  - SCOTTI (Nicola de Maio) does, but unsampled cases are more like an environmental reservoir, convergence issues, hard to use
  - phybreak (Klinkenberg) does, but no unsampled cases. All in R and easy to use.

Thank you. Questions?

- Yuanwei Xu (Birmingham)
- Xavier Didelot (Warwick)
- Christophe Fraser (Oxford)
- Jennifer Gardy (now at Gates)
- Vegard Eldholm (Norway)