

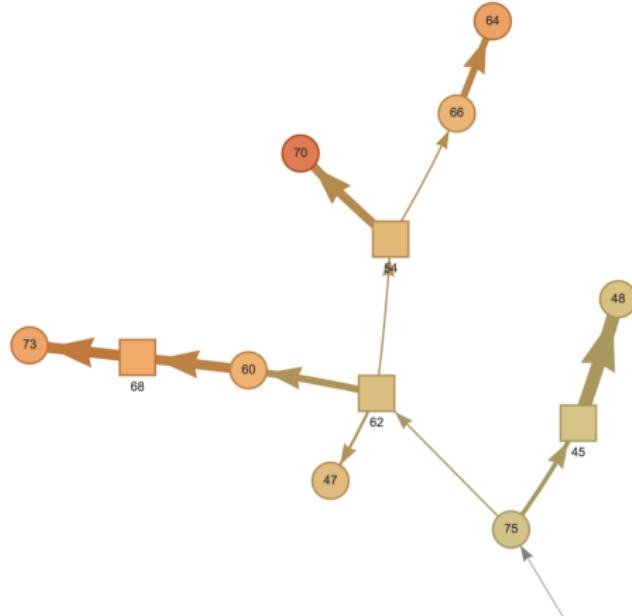
THE MATH AND MOTIVES BEHIND TRANS PHYLO

Caroline Colijn

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.



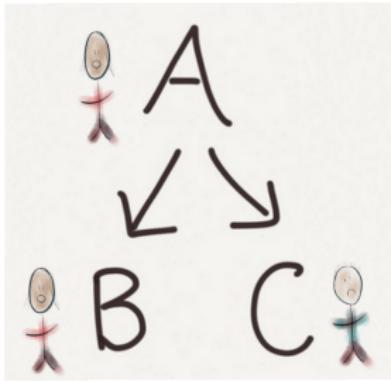
TransPhylo: Use phylogeny to understand transmission

INFORMATION IN TRANSMISSION TREES

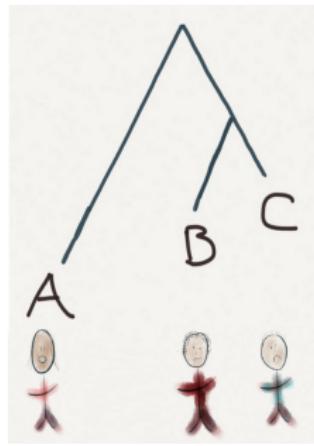
- ▶ who is infecting whom?
- ▶ where is transmission happening?
- ▶ are there variables associated with individuals who have infected others?
- ▶ how soon is transmission happening (after infection)?
- ▶ how long is it taking to sample individuals, after infection?

This kind of information can inform interventions.

EXAMPLE: TRANSMISSION TREE AND PHYLOGENY



A infects B and C



Phylogeny

PURPOSE OF TRANSPHYLO

- ▶ TransPhylo is a software package that uses phylogenetic trees from pathogen sequences to infer transmission trees.
- ▶ We designed it to infer who infected whom and when in an outbreak scenario where only partial data is available.
- ▶ Phylogenetic trees show how different pathogen samples are related to each other, but they don't directly show who infected whom.
- ▶ In our phylogenetic trees, a pathogen sample from an individual is represented as a tip, and the interior nodes represent common ancestors of the samples.
- ▶ In contrast, a transmission tree explicitly shows the paths of infection between individuals. Each node in a transmission tree represents an infected individual, and the branches represent transmission events from one individual to another.

OVERVIEW OF TRANSPHYLO

TransPhylo seeks to infer a posterior collection of transmission trees, given a timed phylogenetic tree, which includes the times when each (sampled) individual was sampled.

It uses a probabilistic model that accounts for within-host diversity in a coalescent model (which describes the relatedness among different pathogen lineages within a host), transmission, and sampling.

TransPhylo accounts for these facts:

- ▶ hosts can have diverse pathogen populations in them
- ▶ not all cases are sampled
- ▶ we know something about the generation time and times to sampling; these are informative about who infected whom

OVERVIEW OF TRANSPHYLO, CONTINUED

TransPhylo does not account for: multiple initial diversity (bottleneck > 1), multiple samples per host, phylogenetic diversity (*).

TransPhylo uses a Monte Carlo Markov Chain (MCMC) method to sample from the posterior distribution of transmission trees, given the phylogenetic tree and sampling times.

This allows it to infer a collection of transmission trees consistent with the data.

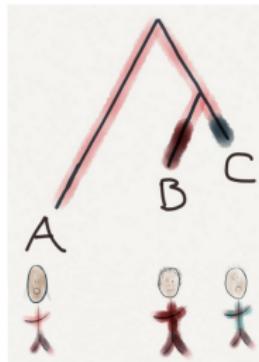
It captures uncertainty in transmission events.

HOW IT WORKS: 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.

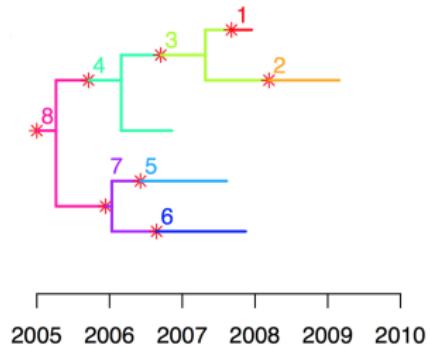


Colour: which host a lineage is in

Each admissible colouring corresponds to a transmission tree.

WHAT IS AN ADMISSIBLE COLOURING?

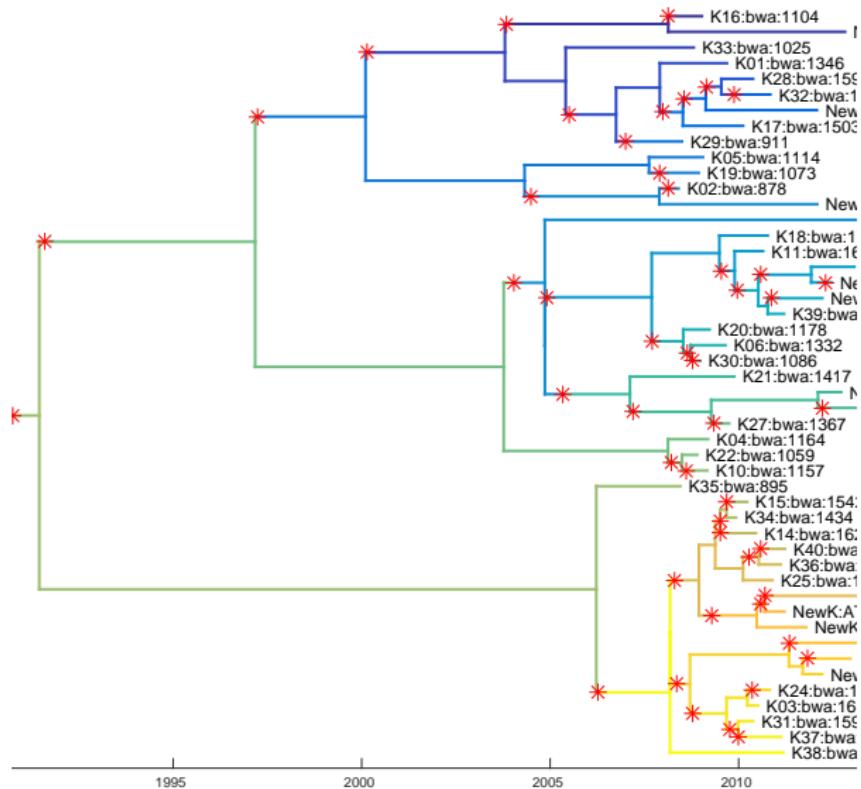
- ▶ Each host has a colour
- ▶ Not all hosts have to be sampled
- ▶ Each lineage is in one host at each time (one colour)
- ▶ Colours can't be broken up (each colour must be continuous on the tree)



ADMISSIBLE COLOURING:

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

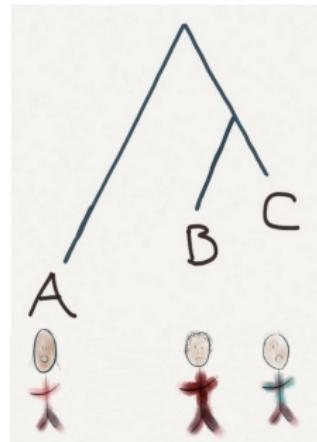
A VALID COLOURING FOR A TB OUTBREAK IN KELOWNA, BC



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. Stage A – make a timed phylogeny (later, we may use more than one). Stage B – do the following (MCMC):

1. Layer transmission events (transmission tree T) on top of it.
This is an example of augmentation.
2. Compute the transmission tree likelihood
3. Accept or reject the state according to the MCMC acceptance probability

Bayes' theorem gives us (with L = likelihood)

$$\begin{aligned} L(Epi, N_{eg}, T | G) &\propto L(G | Epi, N_{eg}, T) L(Epi, N_{eg}, T) \\ &= L(G | N_{eg}, T) L(T | Epi) L(Epi) L(N_{eg}) \end{aligned}$$

WE USE THE COLOURING TO COMPUTE THE LIKELIHOOD

TRANSMISSION

- ▶ Epi : epidemiological parameters defining the transmission process
- ▶ T : transmission tree – who infected whom, and when
- ▶ Colour changes are transmission events – these define T
- ▶ Likelihood: from a 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_{eg}

DERIVATION OF THE DECOMPOSITION

Recall conditional probability:

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

$$\begin{aligned} L(Epi, N_{eg}, T | G) &= L(Epi, T, G | N_{eg})L(N_{eg})/L(G) \\ &\propto L(Epi, T, G | N_{eg})L(N_{eg}) \\ &= L(T, G | N_{eg}, Epi)L(Epi)L(N_{eg}) \\ &\propto L(G | T, N_{eg}, Epi)L(T | N_{eg}, Epi)L(Epi)L(N_{eg}) \end{aligned}$$

G is independent of Epi if you know T . T is independent of N_{eg} if you know Epi . Last two terms are priors. We have:

$$L(Epi, N_{eg}, T | G) \propto L(T | Epi)L(G | N_{eg}, T)Pr(Epi)Pr(N_{eg})$$

THE LIKELIHOOD HAS TWO PARTS: TRANSMISSION TREE; LITTLE MINI-PHYLOGENIES

$$L(Epi, N_{eg}, T|G) \propto L(T|Epi)L(G|N_{eg}, T)Pr(Epi)Pr(N_{eg})$$

L (Transmissions given epi parameters):

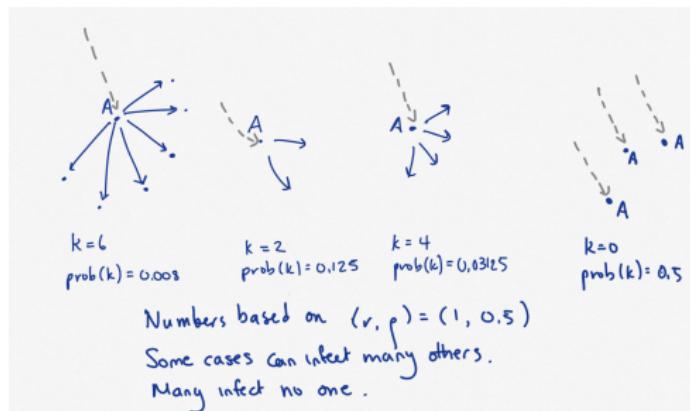
- ▶ Epidemic model for the system: latency, time to infection, time to sampling
- ▶ Finite time due to study end (or the present): this modifies the distribution secondary cases depending on infection time and the sampling probability

L (Phylogeny|Transmission events, coalescent parameter) :

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

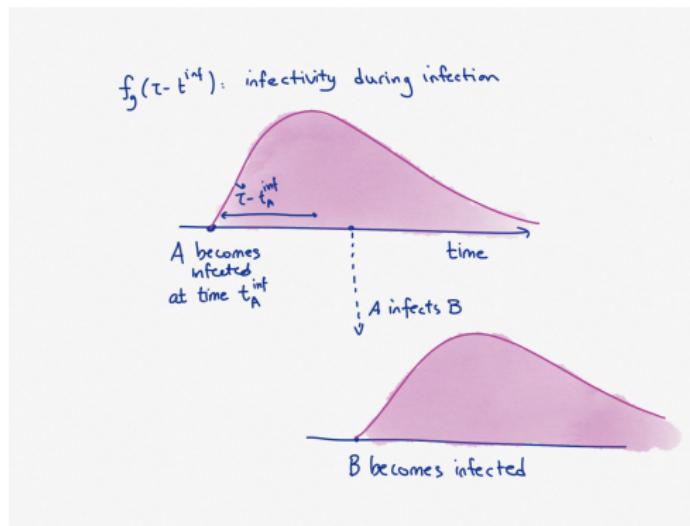
THE EPIDEMIOLOGICAL MODEL (“EPI”) - 1: HOW MANY SECONDARY INFECTIONS?

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



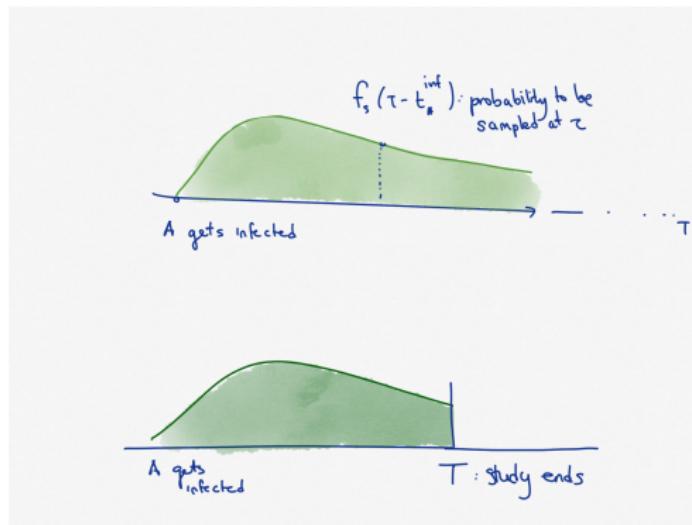
TIME TO INFECT OTHERS

- ▶ 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 ($t \rightarrow -\infty$), and $p_0(-\infty) = p_0^*$. 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. We don't know k so we have to sum over all the (mutually exclusive) possibilities.

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 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

$$\begin{aligned} p_0(t) &= (1 - p_s(t)) \sum_{k=0}^{\infty} p(k) \left[\int_t^{\infty} f_g(\tau_j - t) p_0(\tau_j) d\tau_j \right]^k \\ &= (1 - p_s(t)) \sum_{k=0}^{\infty} p(k) \bar{p}_0(t)^k \quad (2) \end{aligned}$$

WE CAN USE THE PROBABILITY GENERATING FUNCTION

- ▶ Probability generating functions are sums just 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 not, i contributes a $1 - \pi$ (π is the overall sampling probability)
- ▶ If so, use likelihood for the time of sampling for case i
- ▶ How many sampled infectees did i have? Use the probability that i had d_0 sampled descendants, ie $p(d_0, t)$
- ▶ What times did i have these descendants? Use $\prod_{j=1}^{d_0}$ (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_{inf}^i and t_i^s are times of infection, sampling. Then:

$$L(T|Epi) = \prod_{i=1}^n (1 - \pi)^{1-s_i} (\pi f_s(t_i^s - t_{\text{inf}}^i))^{s_i} p(d_0^i, t_{\text{inf}}^i) \prod_{j=1}^{d_0^i} f_g(t_{\text{inf}}^j - t_{\text{inf}}^i)$$

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 $L(\text{Trans}|\text{Epi})$ 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

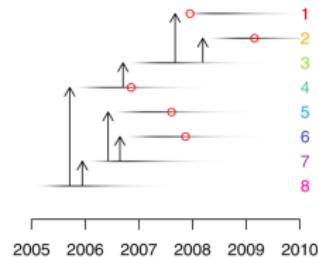
This approach takes in a fixed phylogenetic tree and priors, and produces:

- 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?

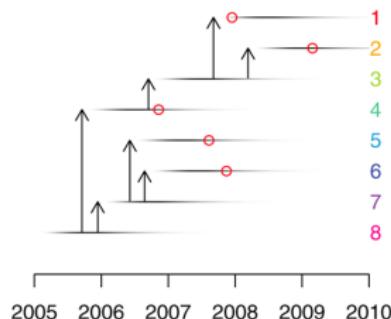
- ▶ A timed phylogenetic tree (or a posterior collection of them)
- ▶ Sampling dates for the tips (ie the isolates)
- ▶ A prior for the time between getting infected and infecting someone else
- ▶ A prior for the time between getting infected and getting sampled
- ▶ A prior for the overall probability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

WHAT DOES TRANS PHYLO PRODUCE?

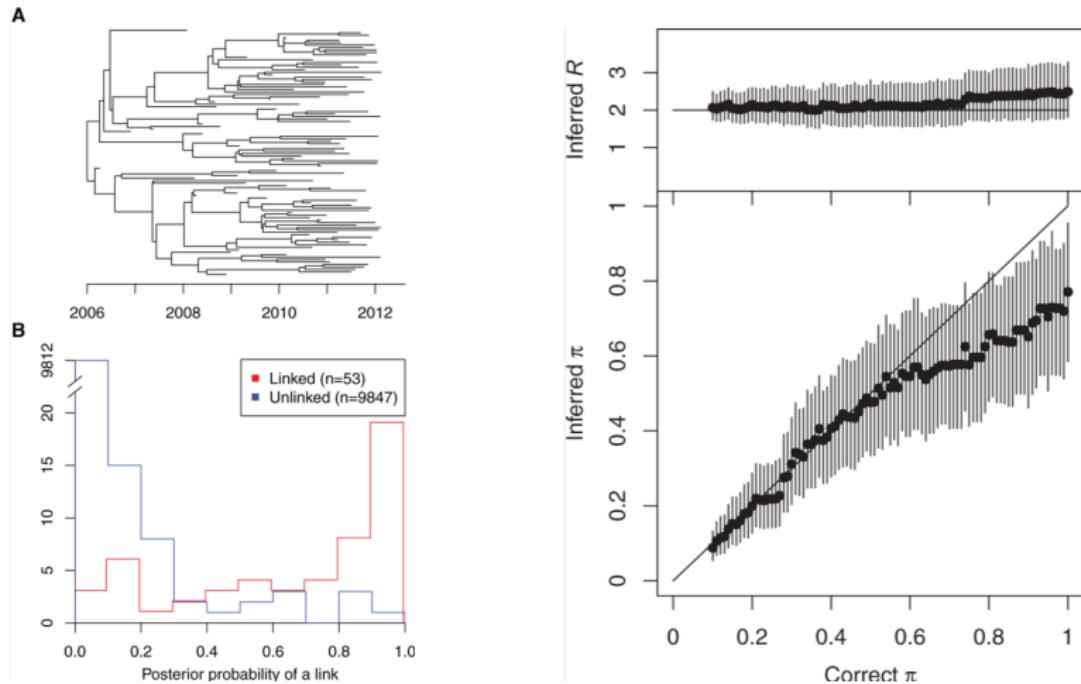
Formally, TransPhylo estimates 3 key parameters: the mean of the offspring distribution (R_0 , in epidemiology), the in-host effective population size, and the sampling fraction.

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

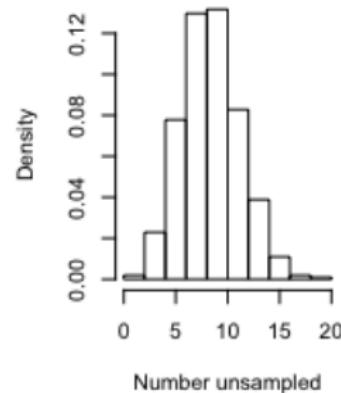
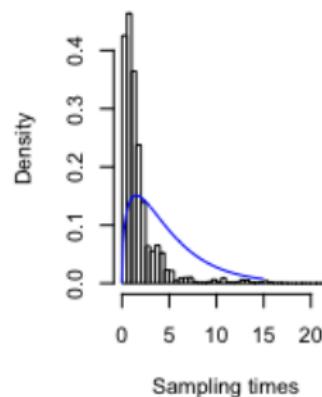
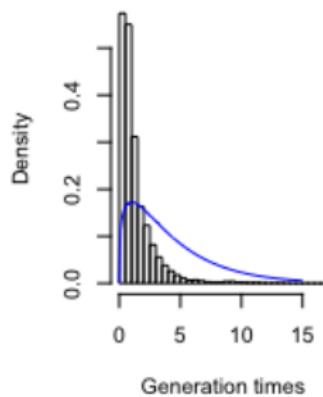
A 13-YEAR TB OUTBREAK IN HAMBURG

- ▶ Outbreak of 86 tuberculosis cases over 13 years, 1997-2010.
Roetzer et al 2013, *Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study*, PLOS Medicine 2013,
<https://pubmed.ncbi.nlm.nih.gov/23424287/>
- ▶ Active case finding among contacts of cases
- ▶ Cases also identified for reasons other than TB infection
- ▶ The outbreak nearly ended shortly before 2006 but spread to a different city
- ▶ TB has potentially long latency, so time from infection to infecting others, and to sampling, are variable (and long)
- ▶ Note that only people with active TB disease can be in the data, or infect anyone in the data

TIME TO INFECTION, TIME TO SAMPLING, NUMBER UNSAMPLED

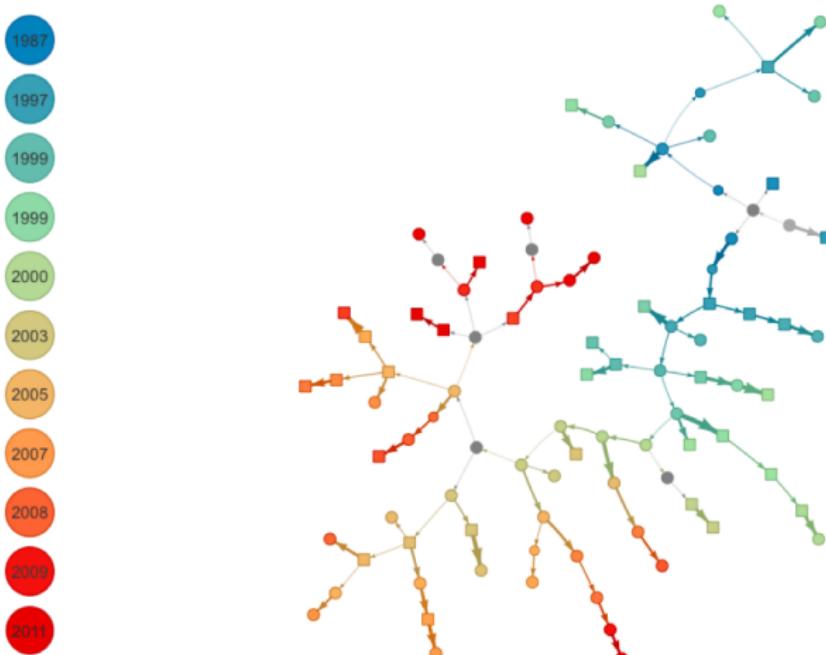
Cases infected someone within 2 years (80%) *among transmitting cases*. 75% sampled in 2 years.

It is likely that some cases were unsampled.



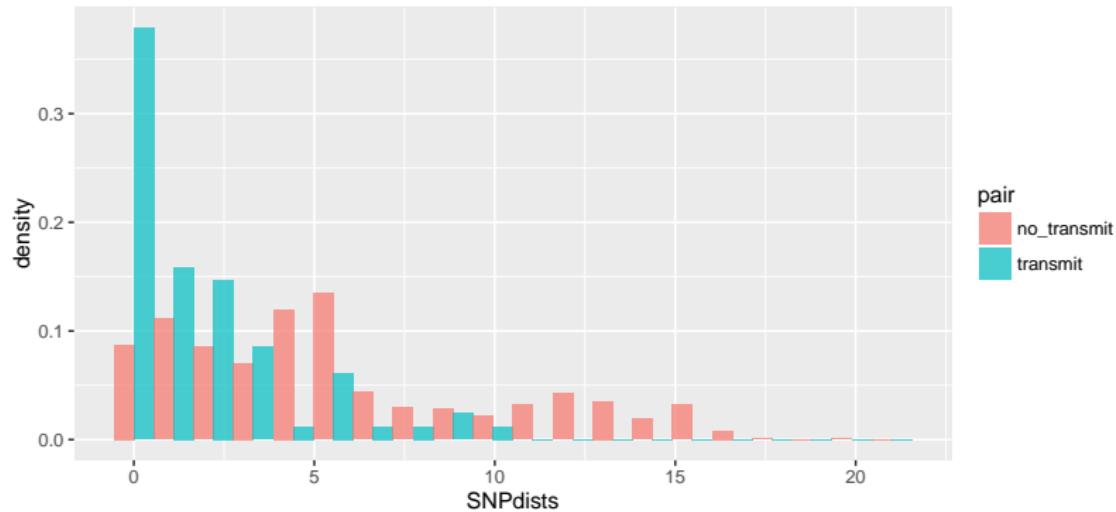
Prior in blue

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?

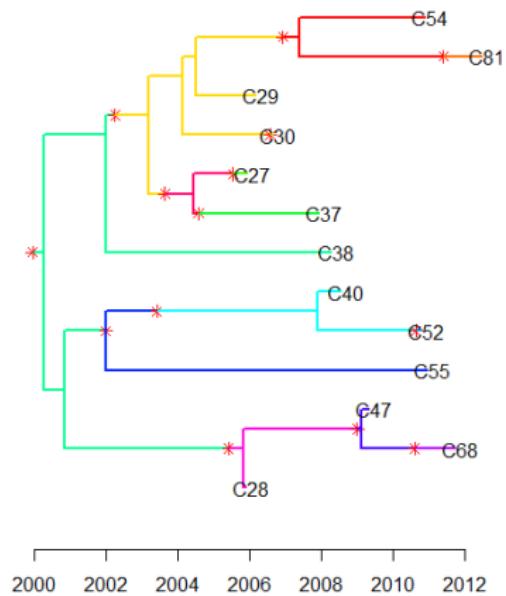
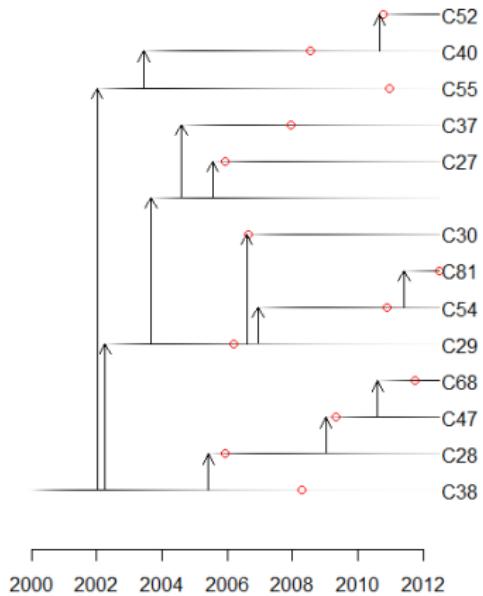
TransPhylo combines the priors for generation and sampling time, plus the genetic data, to give posterior times of infection for each case.

Data:

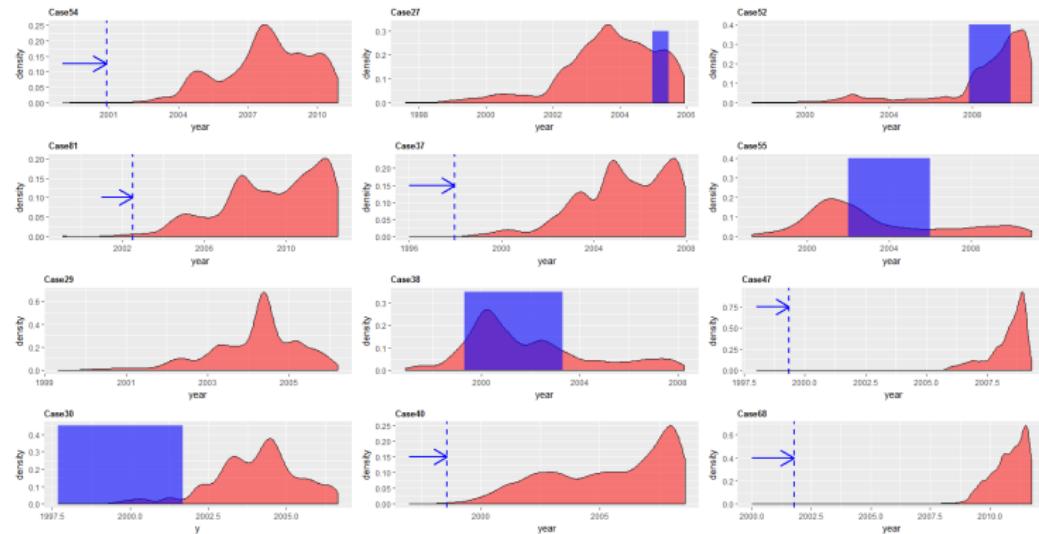
- ▶ Cluster of closely-related cases in Norway
- ▶ Cases occur among people immigrating to Norway
- ▶ It is often assumed that they were infected before arriving
- ▶ But genomic data show signs of recent transmission

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

POSTERIOR TRANSMISSION TREE EXAMPLE



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:
 - ▶ pairwise methods do not limit things enough
 - ▶ methods that assume that branching events in the phylogeny are the same as transmission events limit things too much
- ▶ Good treatment of the sampling process
- ▶ Flexible epidemiological model
- ▶ Even if there is a lot of uncertainty in who infected whom, useful quantities can be extracted:
 - ▶ Timing of transmission
 - ▶ Areas of the phylogeny with many unsampled cases
 - ▶ Who is a “plausible transmitter”? (e.g. infects someone in > 50% of the posterior samples?)

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.

APPLICATIONS: MOLDOVA TB

- ▶ Prospective, genomic analysis conducted on all culture-positive TB cases in the Republic of Moldova in 2018 and 2019.
- ▶ Phylogenetic methods: identify putative transmission clusters.
- ▶ Spatial and demographic data: describe local transmission of tuberculosis.
- ▶ 2,236 participants: 779 (36%) had MDR-TB.
 - ▶ 50% of those with MDR-TB had never been treated previously for TB.
 - ▶ 92% of multidrug-resistant *M. tuberculosis* strains belonged to putative transmission clusters.
- ▶ Phylogenetic reconstruction identified three large clades, comprised nearly uniformly of MDR-TB.
- ▶ Spatial and temporal proximity between pairs of cases within a cluster are associated with greater genomic similarity.
- ▶ BUT: Only two years of sampling – short for TB. Limits transmission analysis.

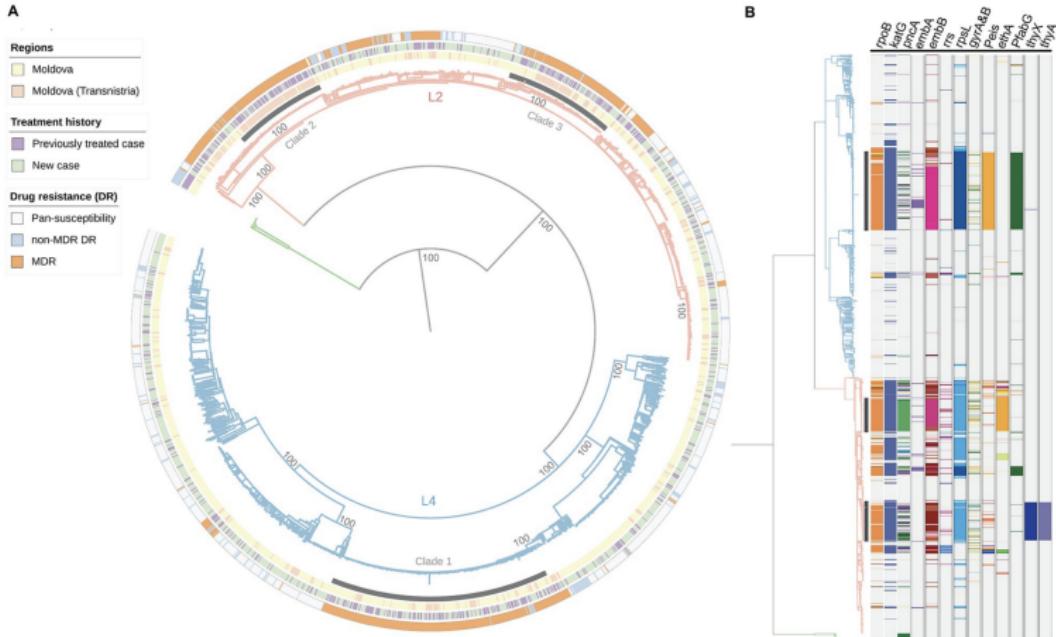


Fig 2. (A) M-L phylogeny of 1,834 Moldova *M. tuberculosis* isolates based on 43,284 variable sites. The outer bands represent the in silico drug-resistant profiles, treatment history of participant and the region where the isolates were sampled from. The tree is rooted to *Mycobacterium bovis* (branch in green). L2 denotes lineage 2 (light orange) and L4 lineage 4 (light blue). Three major clades from the Ural/ lineage 4.2.1 (clade 1) and Beijing/lineage 2.2.1 (clades 2 to 3) are shaded. The main nodes of the tree have 100% bootstrap support. (B) Phylogenetic distribution of resistance-related genotypes. The columns depict loci associated with drug resistance. "P" followed by a subscription of gene name indicates the promotor region. Colored bands of each column represent different polymorphisms. DR, drug resistance; MDR, multidrug resistance; MDR-TB, multidrug-resistant tuberculosis; M-L, maximum-likelihood.

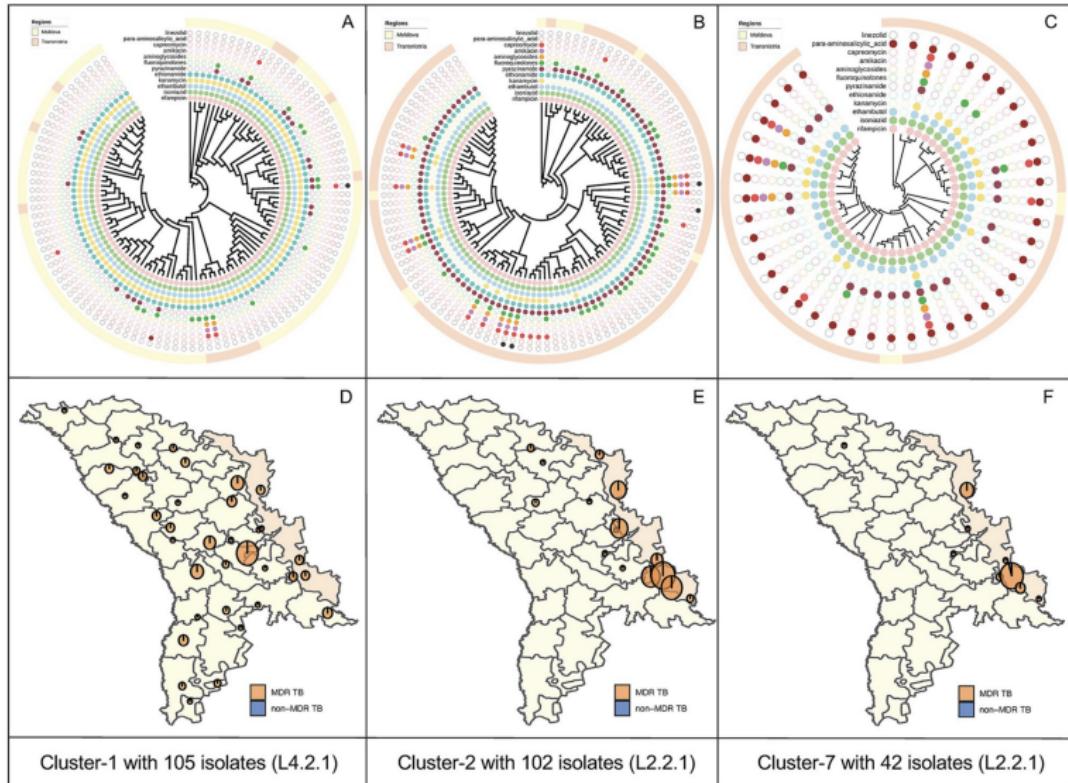


Fig 3. (A–C) Tree visualizations for 3 large putative transmission clusters ($N \geq 10$ isolates), each showing the location of cases in either the Moldova or Transnistria regions along with resistance/susceptibility to 12 anti-TB drugs, as identified by in silico prediction. **(D, E, F)** Spatial distribution of 3 largest clusters (Cluster 1, 2, and 7) in the Ural/Lineage 4.2.1 and Beijing/lineage 2.2.1 clades. The map data were extracted from the GADM database (www.gadm.org/download_country.html). MDR-TB, multidrug-resistant tuberculosis; TB, tuberculosis.

MOLDOVA SUMMARY

- ▶ We reconstructed transmission networks in the 35 broad clusters using the multtree TransPhylo approach and inferred 194 person–person transmission events.
- ▶ Short study period: limited opportunities to capture transmission chains and pairs. 338/1000 clustered isolates were predicted to be involved in transmission events in at least half the posterior transmission trees.
- ▶ But this supports recent, local transmission between sampled individuals in the region.
- ▶ No significant factors were associated with inclusion in these person-to-person transmission events compared to other clustered person-to-person pairs.
- ▶ Some evidence for an increased likelihood of transmission linkage between hosts in the Transnistria region compared to the rest of Moldova (OR 1.42, P = 0.02).

APPLICATION: RESISTANCE AND TRANSMISSION IN TB IN SOUTH AFRICA



The Lancet Microbe
Volume 4, Issue 7, July 2023, Pages e506-e515



Articles

Effect of compensatory evolution in the emergence and transmission of rifampicin-resistant *Mycobacterium tuberculosis* in Cape Town, South Africa: a genomic epidemiology study

[Galo A Goig PhD^{a,b}](#)     , [Fabrizio Menardo PhD^c](#) , [Zubeida Salaam-Dreyer PhD^d](#) , [Anzaan Dippenaar PhD^f](#) , [Elizabeth M Streicher PhD^{g,h}](#) , [Johnny Daniels BAⁱ](#) , [Anja Reuter MPH^j](#) , [Sonia Borrell PhD^{a,b}](#) , [Miriam Reinhard^{a,b}](#) , [Anna Doetsch MSc^{a,b}](#) , [Christian Beisel PhD^j](#) , [Prof Robin M Warren PhD^{g,h}](#) , [Helen Cox PhD^{d,e}](#) † , [Prof Sébastien Gagneux PhD^{a,b}](#) †

Show more ▾

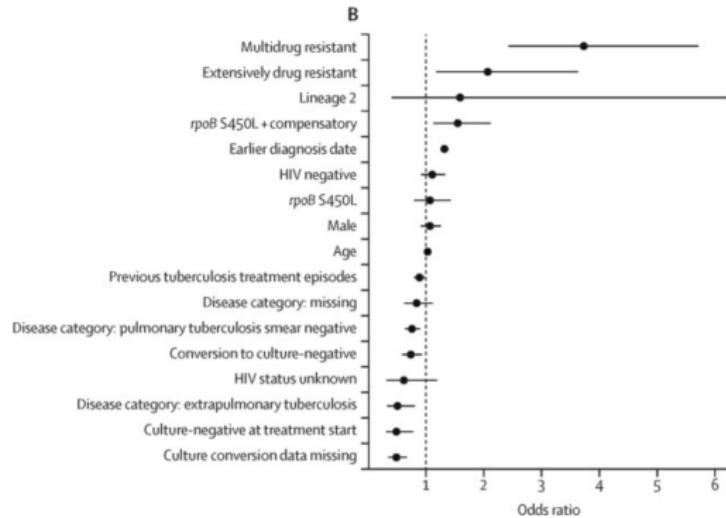
BRIEF OVERVIEW

- ▶ What's the role of compensatory mutations in TB transmission?
 - ▶ Compensatory mutations: mutations that compensate for a fitness cost that comes with resistance-conferring mutations
- ▶ Genomic epidemiology study: 2161 people w multi-drug resistance or rifampicin mono-resistance. 1168 sequences.
- ▶ Time period Jan 2008- Dec 2017
- ▶ Compensatory mutations were associated with smear-positive pulmonary disease, and a higher number of resistance mutations
- ▶ They used TransPhylo to reconstruct transmission.
- ▶ They looked at factors associated with being a (plausible) transmitter

TRANSMISSION ANALYSIS IN THE SOUTH AFRICA PAPER (GOIG ET AL)

- ▶ TransPhylo, separately for each cluster, defined with time to MRCA \leq 15 years; 1M iterations. Used 5000 transmission trees from the last half of the MCMC.
- ▶ Transmitter: inferred to have infected at least one other in at least half of the posterior transmission trees.
- ▶ Due to right truncation, cannot use most recent data: biased.
- ▶ 182 of 838 individuals pre-2016 were identified as “transmitters”

FACTORS ASSOCIATED WITH TRANSMITTING



[Download : Download high-res image \(577KB\)](#)

[Download : Download full-size image](#)

Figure 4. Factors associated with transmission of multidrug-resistant and rifampicin-resistant tuberculosis

(A) Bar plots with counts and proportion of transmitters among groups with different *rpoB* mutations. (B) Phylogenetic multivariable logistic regression of factors associated with being a transmitter of multidrug-resistant and rifampicin-resistant tuberculosis.

THANK YOU. QUESTIONS?

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



SIMON FRASER UNIVERSITY
ENGAGING THE WORLD



Engineering and Physical Sciences
Research Council

