Taking a BREATH (Bayesian Reconstruction and Evolutionary Analysis of Transmission Histories) to simultaneously infer phylogenetic and transmission trees for partially sampled outbreaks

Caroline Colijn[1], Matthew Hall[2], and Remco Bouckaert[3]

[1]Department of Mathematics, Simon Fraser University, Vancouver, Canada; ccolijn@sfu.ca

[2]Nuffield Department of Medicine, Oxford University, Oxford, United Kingdom

[3]School of Computer Science, University of Auckland, Auckland, New Zealand

July 10, 2024

Abstract

We introduce and apply Bayesian Reconstruction and Evolutionary Analysis of Transmission Histories (BREATH), a method to simultaneously construct phylogenetic trees and transmission trees using sequence data for a person-to-person outbreak. BREATH's transmission process that accounts for a flexible natural history of infection (including a latent period if desired) and a separate process for sampling. It allows for unsampled individuals and for individuals to have diverse within-host infections. BREATH also accounts for the fact that an outbreak may still be ongoing at the time of analysis, using a recurrent events approach to account for right truncation. We perform a simulation study to verify our implementation, and apply BREATH to a previously-described 13-year outbreak of tuberculosis. We find that using a transmission process to inform the phylogenetic reconstruction results in better resolution of the phylogeny (in topology, branch length and tree height) and a more precise estimate of the time of origin of the outbreak. Considerable uncertainty remains about transmission events in the outbreak, but our reconstructed transmission network resolves two major waves of transmission consistent with the previously-described epidemiology, estimates the numbers of unsampled individuals, and describes some high-probability transmission pairs. An open source implementation of BREATH is available from `https://github.com/rbouckaert/transmission` as the `BREATH` package to BEAST 2.

1. Introduction

Whole-genome sequence (WGS) data are increasingly used in infectious diseases surveillance for both human and animal populations. Declines in the cost of whole-genome sequencing, and the utility of pathogen sequence data to support epidemiological investigations, have made WGS data appealing for a range of applications. In particular, WGS data can be used to refine knowledge of transmission clusters, to identify links between previously un-linked epidemiological clusters, to characterize the timing of transmission and to identify likely transmission pairs or refute putitative ones [24, 30]. These can, in turn, yield information about the host- and strain-related risk factors for onward transmission [31].

However, even with WGS data, considerable uncertainty can remain about who infected whom. Multiple pairs of individuals may have very closely-related or even identical pathogen genomes [12, 19]. Particularly for chronic infections such as tuberculosis, hosts may harbour substantial pathogen diversity. This complicates the relationship between pathogen sequences (and their similarity between hosts) and whether one host infected another. The timing of symptoms or case detection is not very informative about the infection time for some diseases, including chronic infections such as tuberculosis [14]. The existence of unobserved, or observed but unsequenced, hosts is likely in most settings and complicates the relationship between WGS data and who infected whom. These factors raise substantial challenges in inferring transmission patterns, but sequence data remain a potentially valuable source of information about transmission, as sequences encode variation that is informative about ancestry.

A range of methods have been developed to reconstruct who infected whom – the transmission tree – using pathogen WGS data, and accounting for this complexity and the inherent uncertainties [27, 29]. These include (to name a few) outbreaker2 [18], which integrates epidemiological and genetic data and accounts for unsampled hosts, but does not account for within-host diversity or the shared evolution of the sequences (e.g. through a phylogenetic tree); TransPhylo [16, 25, 33], which accounts for within-host diversity and unsampled hosts but requires a fixed timed phylogenetic tree as an input; BEASTLIER [11] and phybreak [17], which jointly estimate the phylogenetic tree and transmission tree, accounting for within-host diversity, but do not allow for unsampled hosts.

Phylogenetic trees are a key tool to analyze sequence data, because they are constructed with molecular evolutionary models that allow for rate heterogeneity (both between types of base substitutions and genomic positions), molecular clock estimation, consideration of genome content, and other complexities [9, 5, 13, 20]. Accordingly, it is an advantage to use phylogenetic trees in reconstructing transmission from WGS data, compared for example to using single nucleotide polymorphism (SNP) distances. SNP distances do not account for variations in the evolutionary rate or an evolutionary model. However, at the short time scales of person-to-person transmission, we likely do not have a large number of genetic polymorphisms with which to reconstruct a single phylogenetic tree with high confidence [19]. For this reason, TransPhylo has been adapted to allow for simultaneous inference on multiple input phylogenies to account for phylogenetic uncertainty [23]. This is time-consuming, and the sample of phylogenies cannot really reflect the phylogenetic uncertainty. In addition, since the phylogeny is likely to be quite uncertain, it is particularly important to inform phylogenetic reconstruction with prior knowledge about the process that generated the data, namely a transmission process. This cannot be done if the phylogenetic trees are fixed prior to transmission analysis.

Here we build a method called Bayesian Reconstruction and Evolutionary Analysis of Transmission Histories (BREATH) which jointly estimates transmission events and the timed phylogenetic tree, allowing for both unsampled hosts and within-host diversity. This allows epidemiological information about the transmission process (the likely time between infection and infecting others; likely time from infection to sampling; the fact that the pathogen is spreading via person-to-person transmission, or farm-to-farm), to inform the phylogenetic tree. We use a likelihood based on intensity functions and we account for the fact that only individuals who are "ancestral to the sample" can be known to the model. We implemented BREATH as a package of the BEAST 2 software platform [21], opening access to the wide variety of models and methods already implemented in BEAST 2 and its packages, and to state-of-the-art phylogenetic reconstruction. BREATH avoids requiring the computationally challenging reversible-jump Markov chain Monte Carlo (rjMCMC) framework used by TransPhylo to handle unsampled hosts. Instead, in BREATH, the dimension of the annotated phylogenetic tree remains constant regardless of the number of unsampled hosts.

2. Methods

2.1. Mapping transmission on to a phylogeny

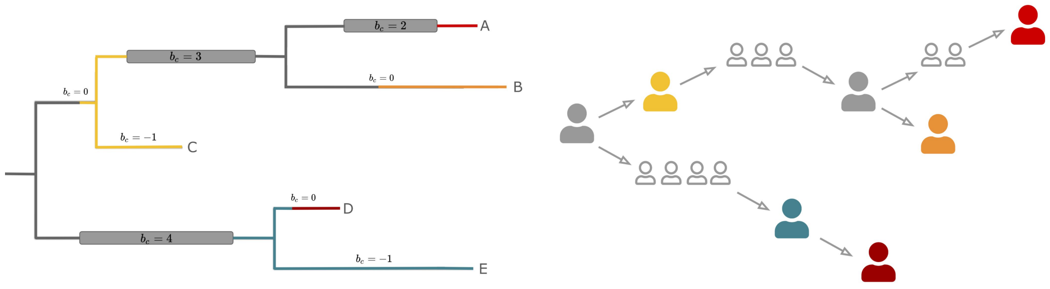
BREATH uses augmentation to annotate a timed phylogeny with a transmission tree, following the same kind of "colouring" approach introduced previously [8, 16, 11]. In previous works, each point on the phylogeny was associated with a specific individual host, whether sampled or unsampled. Note that a "host" may in some applications be an infected location such as a farm rather than an individual human or animal. Here, we will also use a single extra colour that designates a transmission chain of unsampled hosts that occurs on a phylogeny branch. As a result we have two classes of unsampled hosts: "individual" unsampled hosts (IUHs) and members of chains of unsampled hosts. These chains have a duration, making up a subinterval of the branch that they take place on.

We use the terminology "ancestral to the sample" (ATTS) to represent hosts who have at least one sampled host amongst their descendants. (Sampled hosts can be considered technically "ancestral" to themselves for the purpose of our nomenclature.) The sampled descendant host may not be next in the transmission chain, and can be separated from the ATTS host by any number of sampled or unsampled hosts. All hosts of all types in our trees are ATTS; we do not model "unknown unknown" hosts (unsampled hosts with no sampled descendants). "Multiply ancestral to the sample" (MATTS) refers to hosts who are ancestral to at least two hosts that are ATTS. The distinction between IUHs and members of chains is that the former are always MATTS and the latter never are. Sampled hosts may or may not be MATTS.

A MATTS host must colour at least one internal node of the phylogeny, because the common ancestor of the samples descended from both of its child hosts must have existed during its infection. The converse is also true for unsampled hosts: if they colour an internal node then they must be MATTS. Thus the number of IUHs is bounded above by the number of internal nodes of the phylogeny, which for N samples is equal to $N-1$. The number of hosts in unsampled chains, on the other hand, is unbounded. Each chain consists of 1 or more unsampled, non-MATTS hosts and the chain ends when one of these infects a host that is either MATTS or sampled. (This distinction was previously alluded to by [22], in section 3.2 of the appendix.) In this way, we allow for an arbitrary number of unsampled individuals. Each sampled host and IUH is associated with a colour. We require that the subtree corresponding to each of those colours (except the colour corresponding to chains) be continuous – in other words, hosts are not reinfected. Figure 1 illustrates the colouring with a phylogeny and the corresponding transmission tree.

2.2. Bayesian decomposition

Notation: Let G be the genealogy of the pathogen, a timed phylogenetic tree, and T be the transmission tree. The colouring of the nodes of G encodes T: it describes who infected whom and when, and where there are chains of unsampled transmission that are ancestral to the samples. T and G are the two main objects that we want to infer. We also have a number of parameters: θ are the parameters describing the underlying epidemiology, $N_e g$ is the within-host



A. Phylogenetic tree with colouring and block counts

B. Corresponding transmission tree

Figure 1: An example of an annotated phylogeny and corresponding transmission tree. The transmission process starts with an individual unsampled host (IUH; leftmost grey filled individual), in whom the root node resides. U infects C, who is sampled. U also starts a chain of unsampled transmission of length 4. C starts a chain of unsampled transmission of length 3. The chain infects an IUH (rightmost filled grey individual), who infects B, who is sampled. The IUH also infects a chain of length 2, which infects A, who is sampled. In the lower clade, the chain of length 4 infects E, who infects D. They are both sampled. The "block counts" (see *Implementation*) are labelled.

effective population size multiplied by the within-host generation time, and w represents the parameters describing the evolutionary model, including the molecular clock. (The within-host generation time g is that of the demographic process within each host; if T describes spread between organisms it will usually be that of the pathogen, whereas if T instead describes spread between locations such as farms, it may be the serial interval of transmission among individuals in a location.) We assume a strict bottleneck: only one lineage is transmitted at a transmission event. We have sequence data D, and sampling times τ_s. The decomposition is essentially the same as it is in the BEASTLIER model [11], with posterior probability:

$$P(G, T, \theta, N_e g, w | D, \tau_s) = \frac{P(D|G, T, \theta, N_e g, w, \tau_s) P(G, T, \theta, N_e g, w, \tau_s)}{P(D, \tau_s)} \quad (1)$$

where as usual we do not know the denominator but its calculation is unnecessary for inference as it does not vary. Samples from this posterior can be drawn using MCMC.

Conditional on G, we assume the sequence data are independent of T, the parameter $N_e g$, the sampling times and the epidemiological parameters. The first term in the numerator of (1) is

$$P(D|G, T, \theta, N_e g, w, \tau_s) = P(D|G, w) \quad (2)$$

which is the likelihood of the sequence data given the phylogeny and the parameters w describing the evolutionary model and the molecular clock. This likelihood can computed with Felsenstein's pruning algorithm [1] in the usual way.

We decompose the second term in the numerator of (1) further:

$$\begin{aligned} P(G,T,\theta,N_eg,w,\tau_s) &= P(G|T,\theta,N_eg,w,\tau_s)P(T,\theta,N_eg,w,\tau_s) \\ &= P(G|T,\theta,N_eg,w,\tau_s)P(T,\tau_s|\theta,N_eg,w)P(\theta,N_eg,w) \\ &= P(G|T,N_eg,\tau_s)P(T,\tau_s|\theta)Pr(\theta)Pr(N_eg)Pr(w) \end{aligned} \tag{3}$$

The first term, as in [8], describes the likelihood of the genealogy given the transmission tree, within-host parameter and sampling times. This is a product of the likelihoods of smaller (independent) genealogies occurring in different hosts. This is the same as in previous models [8, 16] (see Supplementary Materials). The next term is the likelihood of the transmission tree T and sampling times given the epidemiological parameters. The last three terms describe the priors on the epidemiological and evolutionary parameters.

Putting these ingredients together, we have

$$P(G,T,\theta,N_eg,w|D,\tau_s) \propto P(D|G,w)P(G|T,N_eg,\tau_s)P(T,\tau_s|\theta)Pr(\theta)Pr(N_eg)Pr(w) \tag{4}$$

2.3. Epidemiological likelihood

We wish to find $P(T|\theta,\tau_s)$ from (1). This is the main novelty in BREATH, along with the MCMC moves that enable simultaneous inference of the phylogeny and transmission tree using the colouring. The parameters in θ represent parameters describing the transmission and sampling processes and the time at which observation ends, τ_{end} (we do not infer τ_{end}). We use τ_s to denote all the sampling times.

The likelihood of the transmission tree and sampling times given the epidemiological parameters, $P(T,\tau_s|\theta)$ in (4), can be written recursively. We use τ_i^j to denote j's time of infection, and τ_s^j for j's time of sampling. If j is not sampled, there is no sampling time; in this case we still compute a likelihood, but it describes the likelihood that j was not sampled given j's infection time. Accordingly, set s_j be the sampling event (either sampling at time τ_s^j, or unsampled) for host j. We let $T(k)$ refer to the transmission subtree that begins with host k, and use $j \to k$ to mean "j infected k". We start with the index (root) host r. We have

$$\begin{aligned} L(T,\tau_s|\theta) &= L(T(r),s_r|\theta) \\ &= L(\tau_i^r,s_r,\{T(k),\tau_i^k : r\to k\}|\theta) \\ &= Pr(\tau_i^r)L(s_r,\{T(k),\tau_i^k : r\to k\}|\theta,\tau_i^r) \\ &= Pr(\tau_i^r)L(s_r,\{\tau_i^k : r\to k\}|\theta,\tau_i^r)\prod_{k:r\to k} L(T(k),s_k|\theta,\tau_i^k) \end{aligned} \tag{5}$$

Here, $Pr(\tau_i^r)$ is the prior on the time of infection of the index host. We use an (improper) uniform prior. (Improper priors matter if there is a need to sample from the prior, and when estimating marginal likelihoods through path or nested sampling, but are usually not a problem for sampling the posterior, since the age of the tree is driven by the rest of the prior and by the transmission and sampling intensity functions (see below and *Implementation*).

The index host is the sampled host or IUH associated with the root of the phylogeny. The terms $L(T(k),S_k|\theta,\tau_i^k)$ will in turn become a product of the likelihood of $k's$ events (sampling and infecting others, or not, and the product of the likelihoods $L(T(j),s_j|\theta,\tau_i^j)$ for the js that k infects); infectees may start chains of unsampled transmission. We still need the terms $L(s_r,\{\tau_i^k : r\to k\}|\theta,\tau_i^r)$ for individuals and for unsampled chains. To write these explicitly, we use a recurrent events model, and we treat individual hosts and chains of unsampled transmission separately.

Transmission and sampling

We use a recurrent events approach with two kinds of events: sampling, and transmission to another individual. Each has an intensity function, h^s and h^{tr} respectively. For each,

$$h(t) = \lim_{dt \to 0} \frac{P(\text{event occurs in time } [t, t + dt])}{dt}.$$

In a recurrent events model, the probability density for k events at times $t_1 < t_2 < ... < t_k$ with intensity $h(t)$, in an interval from 0 to t, is [3]

$$\exp\left\{-\int_0^t h(s)ds\right\}\prod_{j=1}^{k} h(t_j).$$

The intensity functions are like hazard functions, except that the event of interest can occur more than once. The expression in the exponential is the probability that the event does not happen in the interval $[0, t]$. It is analogous to a survival function. We will denote these S, with

$$S^{tr}(t) = \exp\left\{-\int_0^t h^{tr}(s)ds\right\} \quad \text{and} \quad S^s(t) = \exp\left\{-\int_0^t h^s(s)ds\right\} \tag{6}$$

for transmission and sampling respectively. For the moment we sample each host a maximum of once, but the model can readily be extended to multiple sampling events.

Since individuals must be infected before infecting others or being sampled, $h^{tr}(t)$ and $h^s(t)$ must be supported only for $t > 0$. We use the gamma shape to allow considerable flexibility in our assumptions about how fast these processes occur: $h^{tr}(t) = C^{tr}$ gamma(t, A^{tr}, B^{tr}), and $h^s(t) = C^s$ gamma(t, A^s, B^s). C^{tr} is the mean number of new infections that an infectious individual is expected to cause over the course of their infection if their infections are not interrupted by the end of the study or by sampling (similar to the basic reproductive number R_0). C^s is the overall sampling probability. A and B refer to the shape and scale parameters respectively. These functions are the building blocks of the likelihood, as they define the contributions from individuals infecting others (or not) and being sampled (or not).

Right truncation

BREATH accounts for the fact that an outbreak might not be over at the time of analysis. This can create bias because faster-occuring events are more likely to be observed. Here, we only have individuals in T if they are either sampled or ATTS: if X and all of X's descendants are not sampled before the end time of our study (τ_{end}), then we never know anything about X at all. This means that our data are right truncated. This is distinct from censoring. Under censoring, we know about an individual, but the event of interest did not occur in our observation window. If the density for observing an event at time t since infection is $f(t)$, but we know that we could only have observed this individual if $t < Y_R$ where Y_R is a right truncation time, then the appropriate contribution to the likelihood for this individual is $f(t|t < Y_R) = \frac{f(t)}{1-S(Y_R)}$ (see [2], Chapter 3).

We begin with the individual host (sampled and IUH). We need to know the appropriate $S(Y_R)$. Individuals are only in the data if they are either sampled or ATTS by the end time τ_{end}. Let the intensity function for the event: "either infects someone ATTS or gets sampled by t" be $h^E(t)$, and let the survival function be S^E.

In a time $(t, t + dt)$, in our model at most one event can happen (dt is very small and rates are finite); either the individual can be sampled, or they can infect someone. If they infect someone, for our event (ATTS) to occur, that person has to eventually become ATTS.

Let p_0 be the probability that an infectee and all of their descendant infections are unobserved – the probability of being an "unknown unknown". We derive p_0 with standard methods (see Supplementary Materials). Once we know p_0, we have

$$h^E = h^s + (1 - p_0)h^{tr}$$

and therefore

$$S^E(Y_R^i) = \exp\left(-\int_0^{Y_R^i}(1-p_0)h^{tr}(t)dt\right)\exp\left(-\int_0^{Y_R^i}h^s(t)dt\right) = S^{tr}(Y_R^i)^{1-p_0}S^s(Y_R^i).$$

Y_R^i , the right truncation time for host i, is the time between i's infection (at τ_i^i) and the end of the observation period, τ_{end}: $Y_R^i = \tau_{\text{end}} - \tau_i^i$.

Likelihood for individual hosts

Recall that we need to build the terms $L(s_k, \{\tau_i^j : k \to j\}|\theta, \tau_i^k)$ to have an explicit likelihood for the transmission tree in (4).

For an individual host k who infects some others, indexed by j, we build this likelihood using the intensity h^{tr} at which k infects others, and the intensity h^s for k being sampled. Let t_i^j be the time interval between k's infection and j's ($t_i^j = \tau_i^j - \tau_i^k$). The contribution to the likelihood that comes from infecting others has a term $h^{tr}(t_i^j)$ for each host j that k infects, along with a survival-like term, $S^{tr}(\tau_{\text{end}} - \tau_i^k)$ or $S^{tr}(t_s^k)$, reflecting the fact that no other infection events happened during the time that k was a potential infector. If sampling prevents onward transmission (for example because hosts are treated effectively or isolated, and no longer infect others), then k stops being a potential infector of others when k is sampled. Otherwise, k could potentially infect others until τ_{end}, though late infections are often very unlikely (as determined by the intensity h^{tr}). Let the length of time when k could infect another be

$$t_e^k = \begin{cases} \tau_{\text{end}} - \tau_i^k & \text{if sampling does not prevent further transmission or k is unsampled} \\ \tau_s^k - \tau_i^k & \text{if sampling prevents onward transmission and k is sampled} \end{cases}$$

The transmission process gives a contribution

$$S^{tr}(t_e^k)\prod_{j:k\to j} h^{tr}(t_i^j)$$

to $L(s_k, \{\tau_i^j : k \to j\}|\theta, \tau_i^k)$: $S^{tr}(t_e^k)$ for no transmission events happening at times other than t_i^j throughout the time when k could infect others, and for each transmission event that does occur, a term h^{tr} for its likelihood. If k does not infect any others, there is no product term.

Similarly, let the relevant time interval for k's sampling be

$$t_s^k = \begin{cases} \tau_s^k - \tau_i^k & \text{if k is sampled} \\ \tau_{\text{end}} - \tau_i^k & \text{otherwise} \end{cases} \tag{7}$$

Throughout, τ values are in "calendar time" and t values denote time intervals. The sampling process has a term with $h^s(t_s^k)S^s(t_s^k)$ if k is sampled, and only $S^s(t_s^k)$ otherwise.

Putting this all together, we have the likelihood for an individual host k's time of sampling (if sampled) and of infecting others:

$$L(s_k, \{\tau_i^j : k \to j\}|\theta, \tau_i^k) = \frac{1}{1 - S^E(T)} h^s(t_s^k)^{\mathbb{1}_{k \text{ sampled}}} S^s(t_s^k) S^{tr}(t_e^k) \prod_{j:k\to j} h^{tr}(t_i^j) \tag{8}$$

This is the main ingredient in (5). The term in the product in (5) breaks down recursively into a collection of terms like this one, for individual hosts, and terms for the chains of unsampled transmission, which we treat next.

Likelihood for unsampled chains of transmission

We now derive the likelihood for the chains of unsampled individuals. Such a chain occurs when an infector infects someone who is not sampled, and so on, until one of two things occurs: either an infectee is sampled, or an infectee is MATTS (multiply ancestral to the sample; see above). That final infectee is not part of the chain even if they are unsampled (in which case their infection must contain at least one internal node of the phylogeny). The rightmost grey host in Figure 1 illustrates this: the chain of length 3 ends because this host is MATTS (infecting both B and another chain). Its infection contains a phylogenetic node. Also, the index case in Figure 1 hosts an internal phylogenetic node, has two lineages that are ultimately ATTS and so is MATTS (ATTS in two ways in this case: through C and the chain). This chain ends when E is infected, because E is sampled. A block has three parameters: the number of hosts n ($n = b_c$, the block count, if $b_c > 0$), the duration of the block, t, and the right truncation time (or, equivalently, the block's start and end times). In the product term $\prod_{k:r\to k} L(T(k), s_k|\theta, \tau_i^k)$ in (5), suppose k is a block that ends when host m is infected. We have

$$\begin{aligned} L(T(k), s_k|\theta, \tau_i^k) &= L(n_k, \tau_{end}|\theta, \tau_i^k)T(m, s_m|\theta, \tau_i^m) \\ &= L(n_k, t_k|\theta)T(m, s_m|\theta, \tau_i^m) \end{aligned} \tag{9}$$

with block duration $t_k = \tau_i^m - \tau_i^k$ and number of hosts n_k. We need $L(n_k, t_k|\theta)$. We drop the subscript k in what follows.

The number of unsampled infections in the block is geometric, because for each infectee X in the chain, there is a probability that X either is sampled or X infects someone who is MATTS. The number of infections in the block is the number of trials that happen before one is successful, which means $n \sim \text{geom}(\rho)$. We need the success probability ρ. We will use the same idea we used for h^E above to find ρ and also to manage the right truncation that occurs because we cannot observe a chain at all unless a MATTS individual is infected by the chain before the observation period ends. We proceed by finding ϕ, the probability that a host will eventually be MATTS, and then ρ, which gives the necessary likelihood:

$$L(n, t|\theta) = \frac{p(n, t)}{1 - P(t > Y_R^j)} \tag{10}$$

where Y_R^j is the block's right truncation time and t is the duration of the block. Details are in the Supplementary Materials.

2.4. Implementation

The transmission tree likelihood, the MCMC proposals, a simulator and some post-processing tools are implemented in the `BREATH` package for BEAST 2 [21].

The BEAST 2 implementation associates three parameters with each branch of the phylogeny: a block count b_c, a block start b_s and a block end b_e. The b_s and b_e define proportions of the branch making up the block length, so $0 \le b_s \le b_e \le 1$ There are three modes of interpretation for block count b_c (see Figure 1):

- $b_c = -1$ represents no infection happening on this branch. Values of block start and block end are ignored in the transmission likelihood. In this case the host (and colour) associated with the phylogenetic nodes at both ends of the branch must be the same. Conversely, if the colours at both ends of a branch are the same, by the continuity (no reinfection) requirement we must have no unsampled hosts on the branch, and $b_c < 0$. If there were reinfection, such that for some hosts, multiple distinct sequences were available, the separate infections could be included as distinct taxa and labelled as if they were distint hosts.

- $b_c = 0$ represents that a single infection took place on the branch, and the host at the top of the branch infects the host at the bottom of the branch, so there are no unsampled hosts in between. MCMC proposals ensure that the block start equals the block end ($b_s = b_e$) for this branch.
- $b_c > 0$ represents the presence of a chain of unsampled hosts and the block start cannot be equal to the block end ($b_s < b_e$).

By initialising b_c to 0 and $b_s = b_e = 0.5$ for each branch, we get a valid starting state.

2.5. Dimensionality

We associate each point in the phylogeny either with a sampled host, an IUH, or a chain of unsampled transmissions. We associate each sampled host or IUH with a colour. (There are a maximum of $2N$ colours for a tree with N taxa: N sampled hosts, up to $N-1$ IUHs and a colour for chains.) Each edge of the phylogeny is given a number b_c encoding the number of transmissions on the branch. Each edge has 2 other parameters (b_s and b_e above). This preserves the dimension of the phylogenic tree object even if the number of unsampled hosts varies. Our representation of chains of unsampled transmission is a mechanism to add unsampled hosts without changing the dimension of the augmented object during the MCMC; the dimension is always $2N-2$ phylogenetic branches and lengths plus the root edge (this allows that the index host, who is always either sampled or an IUH, does not necessarily have a coalescent event at the moment of their infection), and $6N-6$ additional parameters (3 for each edge). Thus the need for the reversible jump MCMC approach of TransPhylo [16] is avoided.

2.6. MCMC proposals

There are two MCMC operators that do proposals for b_c, b_s and b_e: the infection mover and the block operator.

The infection mover, as the name suggests, picks an infection and moves it elsewhere.
• First, it randomly picks two distinct leaf nodes, and retrieves the path between the two nodes. Since leaf nodes represent different hosts, there must be at least one infection along the path, but possibly there can be multiple infections, for example in one or more blocks.
• Next, it uniformly selects an infection from the set of transmissions and removes it from the path by reducing the block count by 1. If b_c becomes 0, b_s and b_e are set to become equal, either by setting b_s to b_e or b_e to b_s, each with 50% probability. If b_c stays > 0, block boundaries are updated as well so that the block length shrinks: with 50% probability b_s is uniformly selected between the original b_s and b_e and otherwise b_e is uniformly selected in that interval.
• Finally, uniformly in the length along the path randomly select a point x. A new infection is added by finding the appropriate branch that contains x and increasing its b_c by 1. If b_c becomes 0, $b_s = b_e = r$ where r is the fraction of x on the branch. If $b_c > 1$, detect whether r is inside the block and if not, adjust b_s or b_e such that r is on the boundary of the block. The Hasting ratio for this operator is 1.

The block operator with 50% probability only moves boundaries, and otherwise removes or adds infections.
• When moving block boundaries, it randomly picks a branch uniformly (not taking branch lengths into account). If $b_c < 0$ it returns without doing anything else, if $b_c = 0$ it uniformly picks a value between 0 and 1 and assigns it to both b_s and b_e. If $b_c > 0$, with 50% probability

it assigns to b_s a new value uniformly chosen between 0 and b_e, and otherwise assigns to b_e a new value between b_s and 1. The Hasting ratio is 1 when only moving block boundaries.

- Otherwise, infections are removed or added (with the remaining 50% probability). First, consider removing an infection. Not all infections can be removed without leaving the transmissions in an invalid state, for example, a single infection on a branch between two sampled hosts cannot be removed. First all infections are determined that can be removed without causing problems, then uniformly one is chosen among those eligible infections (or the proposal fails when no such infection is available) and the move proceeds as follows: When $b_c = 0$, reduce b_c to -1. When $b_c = 1$, reduce b_c to 0 and with 50% probability set b_s to b_e otherwise set b_e to b_s. When $b_c > 1$, reduce b_c by one and leave block b_e to b_s without change. The Hasting ratio is the branch length divided by the tree length times number of eligble nodes to remove.
- Now, consider adding an infection. Infection are added by selecting a branch with probability proportional to the length of the branch. If $b_c < 0$, set b_c to 0 and assign $b_s = b_e = r$ where r uniformly chosen between 0 and 1. If $b_c = 0$, set b_c to 1 and r uniformly chosen between 0 and 1; if $r < b_s$, set $b_s = r$ otherwise $r > b_e$ and we set $b_e = r$. If $b_c > 0$, increaset b_c by 1. The Hasting ratio is 1.

Since there are no parameters involved in any of the block count changing proposals, there is nothing that can be optimised during execution of the MCMC algorithm. For the population size parameter N_eg, we used a standard Bactrian scale operator.

2.7. Parameters for the outbreak analysis

We analyzed a previously-described TB outbreak [6] with 86 detected cases in Germany between 1997 and 2010. Single nucleotide polymorphisms are available in the previous publication along with month and year. We used $A^s = 10.0$, $B^s = 6.5$ and $C^s = 0.75$ for the sampling intensity; $A^{tr} = 10$, $B^{tr} = 8.5$ and $C^{tr} = 2$. These were informed by the analysis of the same outbreak with TransPhylo [16]. We used the bModelTest site model [15] and a strict molecular clock. We used an improper uniform prior for N_eg with a lower bound of 0.1. Other parameters are as specified in the XML file at `https://github.com/rbouckaert/transmission/releases/tag/v0.0.1`.

2.8. Availability

BREATH is available at `https://github.com/rbouckaert/transmission/` with a tutorial at `https://github.com/rbouckaert/transmission/tree/main/doc/tutorial`. The data for the TB outbreak are available in [6] with a demostrative dataset of 40 sequences in the tutorial.

3. Results

3.1. Validation

We performed a well calibrated simulation study [34] to provide some confidence that the implementation is correct. To this end, we first sampled 100 transmission trees with 32 taxa using the `TransmissionTreeSimulator` app. For the sampling intensity, we used a shape A^s of 10, a rate B^s of 5 and a sampling probability C^s of 0.75. For the transmission intensity, we used a shape A^{tr} of 10, rate B^{tr} of 8 and expected infection count C^{tr} of 2. The N_eg parameter size was set at 0.5.

Further, we simulated an alignment on the tree with 500 sites under an HKY model with $\kappa = 2$ and equal frequencies (no gamma rate heterogeneity) and a clock rate of 0.25. Then, we ran

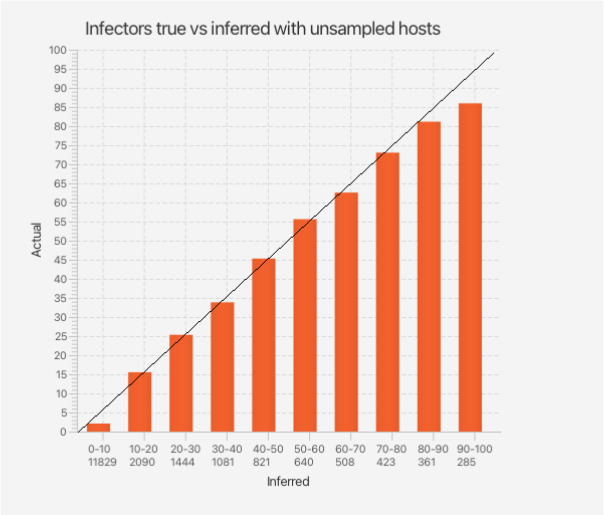


Figure 2: Who-infected-whom inferred probabilities versus true values. The graph is constructed by counting how many inferred probabilities fit in a bin, then calculating how many of the true values are in a certain bin. Each bar represents a 10% sized interval, where the first bar represents predictions in the range 0 to 10%, the second bar ranges from 10% to 20%, etc. The number below the range represents the number of predictions that fit in that range for each of the 100 posterior distributions. For example, there are 284 predictions in the range 90-100% (last column). The size of the bar is the percentage of times the true prediction is in the bin with the associated prediction. Apart from perhaps the very high confidence predictions, all other predictions are as expected.

analyses for each of the 100 transmission trees in BEAST 2 where the tree, block count, block start and end as well as the population size were estimated. We measure how often a true parameter value is in the 95% highest probability density (HPD) interval, which for 100 runs should be in the range 90 to 99 to be acceptable. The coverage of tree length is 92, tree height 90, but the infection count (the total number of infections in a tree) had a coverage of 84, which is outside the acceptable range. This is probably due to the subtle way the model differs from the simulator: the transmission tree likelihood does not condition on the number of taxa, but the trees for the simulation study are selected to have a fixed number of sampled hosts. So, the trees in the simulation study are conditioned on having 32 taxa.[1]

We want to verify that if the model predicts that host A is infected by host B with probability p that the true probability under the model is indeed p. Figure 2 shows the who-infected-whom predicted probabilities and their "true" probabilities (how often events occurred in the simulated ground truth). Note that the size of each bin, corresponding to the number of predictions with the given probability of being infected by a particular host, decreases with increasing probabilities. So, there are many very low confidence predictions, and the higher the confidence, the lower the number of predictions. Ideally, all bars should cross the black x-y line, indicating that predicted probabilities equal actual probabilities. Though most bars are very close to the ideal, the very high prediction bars for 80-90% and 90-100% are just under what they should be. Note that a large majority of the predictions in these categories are for being infected by an unsampled host, which are not usually the predictions of interest. This suggests a slight over-confidence in very high probability posteriors, possibly due to a slight mismatch

[1] Files for the well calibrated simulation study are available from `https://github.com/rbouckaert/transmission/releases/tag/v0.0.1`.

between the simulator and the implemented model.

3.2. Application to TB

We analysed an alignment of 86 samples from a tuberculosis outbreak in Hamburg, Germany, described in [6]. We fixed the intensity function parameters as in the simulation study, and estimated the within-host coalescent parameter. Since we are not interested in estimating the site model (better done for tuberculosis with larger datasets comprising more evolution), we average over reversible models using bModelTest [15]. A strict clock is used for the branch rate model. To see what the impact is of the transmission tree likelihood on the phylogenetic tree, we compared BREATH's results with BICEPS [26], a flexible coalescent prior taking epochs into account.[2]

Site models are indistinguishable across the two scenarios, which is as expected: the tree prior should not impact the way sequences evolve. However, the root height and tree length are somewhat lower in BREATH than in BICEPS: the mean height is 21.8 for BICEPS with 16.2-28.9 95% HPD, and 15.4 for BREATH with 14.6-17.7 95% HPD. These correspond to a time of origin of early 1989 (late 1980 – early 1994) in BICEPS, compared to mid-1995 (early 1993 – March 1996) in the transmission process model. The published description of this outbreak listed a time of origin of 1993-1997 for what they termed the "Hamburg clone" (the main driver of this outbreak) [6].

The mean total tree length (sum of all branch lengths) 217.6 years in the BICEPS model (146.4-304.3 95% HPD) and is lower at 161.3 for BREATH (144.9-178.2 95% HPD). In both the branch length and the tree heights, we find a substantial reduction not only in the value but also in the uncertainty of the estimate. The change in node height estimates is most pronounced near the root, and not so pronounced at internal nodes that are placed lower in the tree. Clade support changes between the two analyses (Figure 3), also showing the tree prior has a substantial impact on the inferred trees.

TB arises from a process of person-to-person transmission with a variable time frame of years from infection to infecting others and to sampling (as reflected in our intensity functions). Including this information, as BREATH does, informs the phylogenetic analysis and results in a phylogenetic tree requiring less overall evolution to explain the data than a coalescent model. For context, Figure 5 shows the CCD0 summary tree [32] for BREATH, coloured by number of transmissions per branch, as well as the DensiTree, which visually illustrates the phylogenetic uncertainty. The method has no problem dealing with the amount of phylogenetic uncertainty, which in this case is considerable.

Figure 4 shows BREATH's inferred graph of who infected whom. Nodes in the graph represent hosts, and edges in the graph represent possible transmissions between hosts. The edges are annotated with probability of infection as inferred under the model. Hosts in tightly-connected parts of this graph have a high probability that they were infected by another sampled host (as opposed to an unsampled host); this probability is shown in parenthesis in Figure 4. These "clusters" are in areas of the phylogenetic tree where branch lengths are short, and sequences have short distances to each other; it is these short branches that likely drive the inference of infection by another sampled hosts.

We also compare BREATH's transmission inferences with those of TransPhylo [16] (see Supplementary Materials). We find that BREATH has a more consistent pattern of genetically similar pairs of sequences corresponding to higher posterior transmission probabilities. BREATH's transmission probabilities are lower overall, and BREATH has fewer high-probability trans-

[2] BEAST 2 XML files for the TB analysis are available from `https://github.com/rbouckaert/transmission/releases/tag/v0.0.1`.

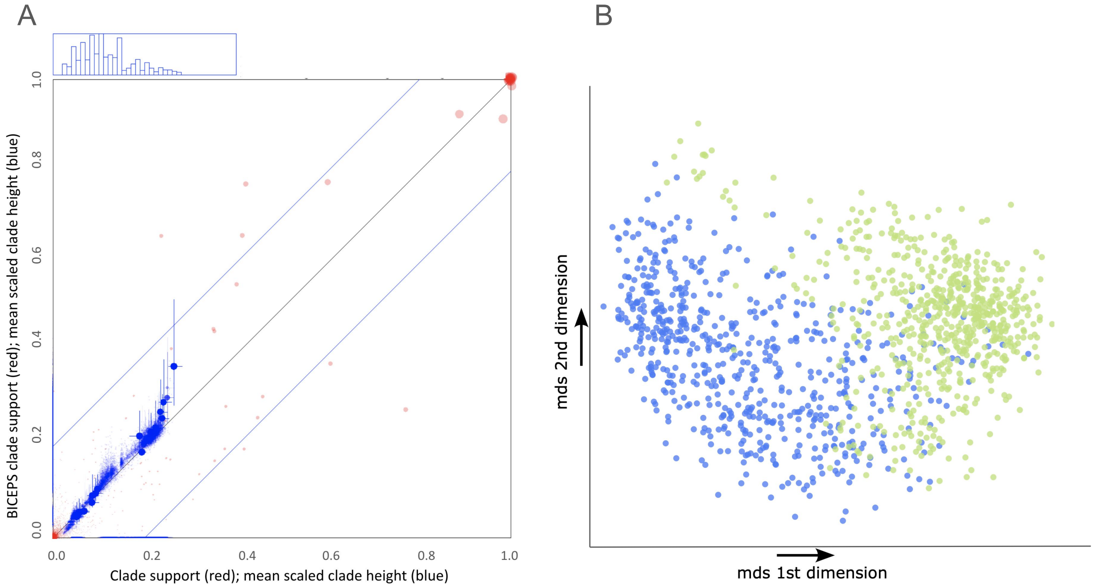


Figure 3: Difference between BICEPS and BREATH. A. Clade support (red dots) running from zero probability (bottom left corner) to 1 (top right corner) and scaled clade heights (blue dots). Dots are placed at the mean of the height of a clade. Both x and y axes are scaled so that the largest tree height (from among both data sets) is at the top right corner. Blue crosses indicate 95% highest probability density (HPD) intervals of a clade's height. Larger dots have larger clade supports. We compare BREATH (horizontal axis) to BICEPS (vertical axis). The diagonal blue lines mark a 20% difference between BREATH and BICEPS: points within these lines have less than 20% difference, points outside have more, and can be considered as evidence of a substantial difference between the two posteriors. B. Multidimensional scaling plot generated by TreeDist package in R [28] showing marked differences between the tree posteriors of the BICEPS and transmission tree analysis.

mission pairs among the samples. Where BREATH's posterior probability for an infection event is high, typically so is TransPhylo's, but the converse is not true (though TransPhylo's highest-probability events do have probability > 0.25 in BREATH). These results suggest that TransPhylo, due to its assumption of a fixed phylogeny, may be over-confident in some transmission pairs. This is because a fixed phylogeny constrains possible sets of transmission events due to the colouring constraint. For example, in Figure 1, if the phylogeny were fixed, then if C infected A, then either C or an unsampled case or chain that C infected would be the only possible infectors for B.

Our results indicate that even with whole-genome sequencing data, if we allow for within-host transmission, unsampled hosts and phylogenetic uncertainty, we are left with considerable uncertainty in who infected whom. Where pairs can be identified, the direction of transmission is often uncertain. Nonetheless, we find some high-probability transmission events, as well as individuals for whom the infector is not known and individuals for whom there is a high probability that the most likely infector was an unsampled individual. We also identify three major sections of the graph: one from 1997 to approximately 2000 (hosts 29, 30, 33; see [6] Table S2; one from approximately then to 2010 (hosts 33 to 83) and one from 2008 (hosts 49 and 51) to 2010. Roetzer et al reported that while there were strong efforts to stop the outbreak, including closing a particular bar in 2006, the outbreak continued in Hamburg and spread to Schleswig-Holstein in 2006. In our posterior trees, in 2002-2004 there were (with 95% posterior probability) fewer than 3 lineages present, so that the outbreak came very close to being contained before being detected again in 2006. A median of 1 of those lineages was in an individual who was never sampled. Overall, we found an average of 20 unsampled (but ATTS) individuals, and the maximum length of an unsampled transmission chain was 2.74 hosts on average (i.e. the average of the maximum-size chain, over the posterior transmission trees, was 2.74).

4. Discussion

BREATH is a Bayesian method that simultaneously constructs phylogenetic and transmission trees, accounting for within-host diversity and allowing for a flexible number of unsampled individuals, and real-time outbreaks. The BREATH tree prior leverages BEAST 2's power in phylogenetic inference: it can be combined with nucleotide, aminoacid and other types of sequence data, various models of evolution of characters along a tree and different clock models.[3] BREATH overcomes a substantial barrier in reconstructing transmission, since previous methods could not account for unsampled hosts while estimating phylogenetic and transmission trees, requiring either simplifying assumptions or a fixed input phylogeny. We applied the method to data from a TB outbreak [6], and found that our transmission process led to more well-resolved phylogenetic trees than standard models. We identified several distinct stages of transmission with only a maximum of 3 concurrent hosts separating them, suggesting that the outbreak came very near to being interrupted in the early 2000s. While there is considerable uncertainty in who infected whom, we identified some high-probability transmission events, distinct subgroups in the outbreak, and found some hosts likely infected by an individual not in the sample. Information about the posterior numbers of secondary infections for each host, which hosts did not have a plausible infector among the sampled individuals, the posterior times between infection and sampling and similar information can readily be extracted from the posterior transmission trees in our model. These are all potentially useful even with uncertainty in individual transmission events.

[3] A tutorial for setting up an analysis with BREATH using BEAUti – the user friendly graphical interface for BEAST 2 – is available at `https://github.com/rbouckaert/transmission/tree/main/doc/tutorial`.

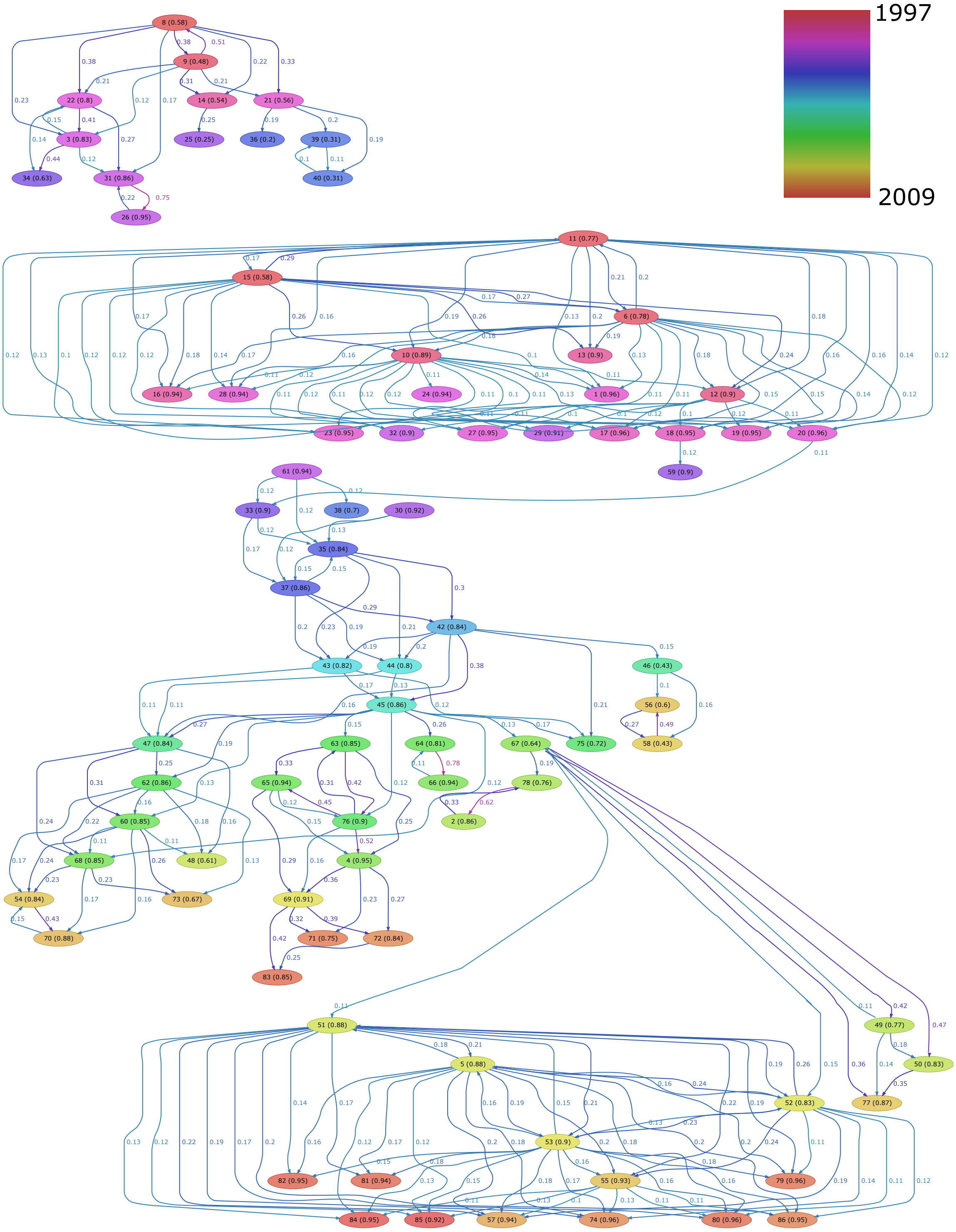


Figure 4: Who-infected-who predicted in TB outbreak. Numbers on branches represent probabilities the host at the tail of the arrow infecting the host at the head. Only the most important predictions (over 10% probability) are shown. Numbers in brackets next to labels are the inferred probabilities a host is infected by another sampled host. Hosts are coloured by time of sampling.

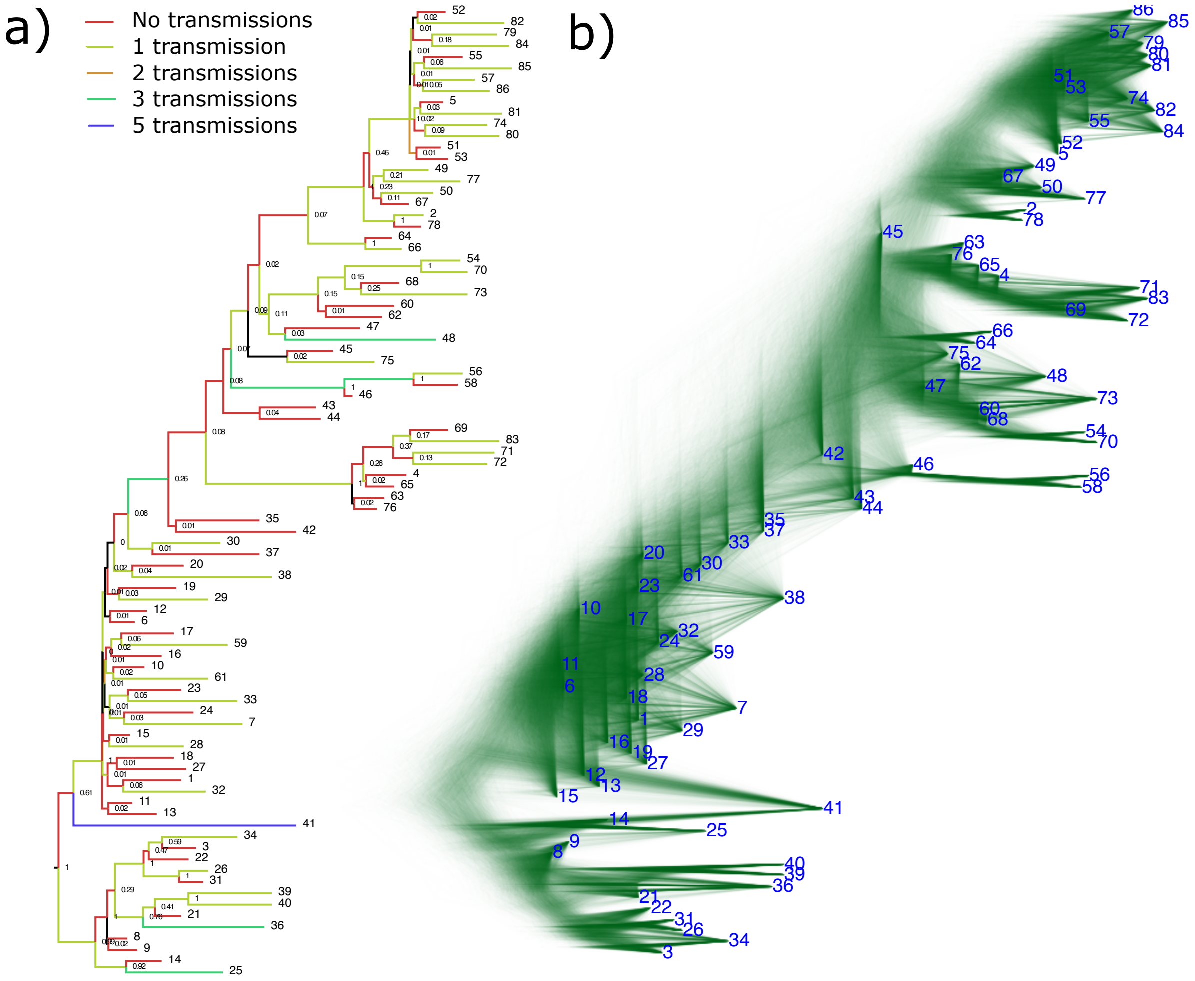


Figure 5: Results from TB outbreak analysis with labels the same as in Figure 4. a) maximum a posteriori tree based with branches coloured by median block count. Numbers on branches are posterior support. Some branches do not have sufficient support for the block count to be reliably estimated and are black. b) DensiTree [4] of the transmission tree analysis, showing considerable uncertainty in both clade support and internal node heights. This is reflected in its entropy of 69.1, which is quite high for a tree with 87 taxa.

BREATH has the limitation that there are multiple interacting parameters that influence branch lengths and therefore also the estimated numbers of unsampled individuals. How these interact, and which part of parameter space enables accurate inference of phylogenies and transmission histories, is still to be determined. As in previous methods, it is not likely to be possible to infer the generation time (from infection to infecting another), the within-host parameter and the numbers of unsampled individuals: without some prior information, there would be no way to distinguish between a branch with many brief inter-transmission intervals or fewer longer ones. Fortunately for most infectious diseases for which sequence data are likely to be available, there is some knowledge of the transmission and sampling timing.

There are several changes that can be made to improve the realism of the model. For example, the model assumes a host is removed from the population when sampled, but it is possible the host infects other hosts after sampling; this is readily modified. It would also be straightforward to relax the assumption of constant within-host $N_e g$ over time and across hosts. An exponential or logistic population trajectory may be more appropriate. Different hosts could also have different coalescent parameters, if there were data to inform these. These changes would straightforward as long as the relevant coalescent likelihood is tractable.

In closely-related outbreaks, the amount of detectable genetic variation is likely to be low, resulting in phylogenetic uncertainty, and uncertainty about who infected whom. For infections for which the sampling time is more informative about infectiousness (than it is for a chronic infection like TB), timing may resolve some of this uncertainty. In any case, it is an advantage of our approach that the phylogeny can be informed by the transmission process; this is this process that generates the data. Phylogenetic trees are used to understand the acquisition vs transmission of antibiotic resistance in TB [7], to resolve times of origin of emerging clades [30] and to infer how pathogens move geographically [10]. In any of these applications, having a better phylogenetic model is helpful even if person-to-person transmissions are not the focus of interest.

BREATH assumes that hosts infect other hosts only once, and a single lineage is transmitted. However, hosts could be infected multiple times, depending on immunity, infectiousness and other factors. Knowledge of multiple infections would require multiple samples per host and/or deep sequencing sufficient to characterize individuals' diverse infections. This raises challenges: the priors or constraints on how a single host's taxa should be placed in the phylogeny depend on whether the host was infected more than once. If reinfections could be assumed to proceed similarly to other infections, BREATH would be readily adaptable to having multiple samples per host, and each augmented phylogeny would indicate, for that posterior sample, whether the host had one infection that diversified (corresponding to all of that hosts' taxa contiguously coloured) or whether there was a reinfection.

The potential benefits of transmission tree inference are not only in human-to-human outbreaks but also in the management of zoonotic infections (such as bovine tuberculosis, foot and mouth disease, avian influenza). Who-infected-whom information is useful in informing policy decisions about whether to quarantine certain areas, cull herds (in case of animal disease), or put movement restrictions in place. In this paper, we modelled hosts as human individuals; hosts could also represent farms affected by a disease outbreak, or non-human hosts.

In some contexts there may be information about the transmissibility and/or susceptibility of different groups of hosts, like children being more or less susceptible or infectious than adults. These aspects could be modelled with different intensity functions for different kinds of hosts. This is straightforward in the likelihood for sampled individuals, but unsampled individuals would be challenging to model. More optimistically, integrating contact data by adding priors/constraints on the colouring of a tree would be straightforward.

The transmission tree likelihood is not particularly computationally intensive compared to the sequence tree likelihood, and BEAST 2 readily scales to many hundreds of sequences. Our model can therefore scale well with larger numbers of samples on the scales for which we would anticipate having densely-sampled sequences. Another advantage is that due to the placement of unsampled transmission chains on branches, a preliminary clustering step – identifying putative transmission clusters for onward analysis, with a SNP cutoff or phylogenetic method – should not be required. In the event of an ongoing outbreak, it would be desirable to have a version that allows re-use of previous analyses when new data becomes available (instead of running the whole analysis from scratch.)

Acknowledgments

This work was funded by the Natural Sciences and Engineering Research Council (Discovery Grant RGPIN-2019-06624) and the Federal Government of Canada's Canada 150 Research Chair program.

References

[1] Joseph Felsenstein. "Evolutionary trees from DNA sequences: a maximum likelihood approach". In: *Journal of molecular evolution* 17 (1981), pp. 368–376.

[2] John P Klein, Melvin L Moeschberger, et al. *Survival analysis: techniques for censored and truncated data.* Vol. 1230. Springer, 2003.

[3] Richard John Cook, Jerald F Lawless, et al. "The statistical analysis of recurrent events". In: (2007).

[4] Remco R Bouckaert. "DensiTree: making sense of sets of phylogenetic trees". In: *Bioinformatics* 26.10 (2010), pp. 1372–1373.

[5] Fredrik Ronquist et al. "MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space". In: *Systematic Biology* 61.3 (Feb. 2012), pp. 539–542. ISSN: 1063-5157. DOI: 10.1093/sysbio/sys029. eprint: https://academic.oup.com/sysbio/article-pdf/61/3/539/24563565/sys029.pdf. URL: https://doi.org/10.1093/sysbio/sys029.

[6] Andreas Roetzer et al. "Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study". In: *PLoS medicine* 10.2 (2013), e1001387.

[7] Nicola Casali et al. "Evolution and transmission of drug-resistant tuberculosis in a Russian population". In: *Nat. Genet.* 46.3 (Mar. 2014), pp. 279–286.

[8] Xavier Didelot, Jennifer Gardy, and Caroline Colijn. "Bayesian inference of infectious disease transmission from whole-genome sequence data". In: *Mol. Biol. Evol.* 31.7 (July 2014), pp. 1869–1879.

[9] Alexei J Drummond and Remco R Bouckaert. *Bayesian evolutionary analysis with BEAST.* Cambridge University Press, 2015.

[10] Vegard Eldholm et al. "Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain". In: *Nat. Commun.* 6 (May 2015), p. 7119.

[11] Matthew Hall, Mark Woolhouse, and Andrew Rambaut. "Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set". In: *PLoS Comput. Biol.* 11.12 (Dec. 2015), e1004613.

[12] Nicola Casali et al. "Whole Genome Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study". en. In: *PLoS Med.* 13.10 (Oct. 2016), e1002137.

[13] Sebastian Höhna et al. "RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language". In: *Systematic biology* 65.4 (2016), pp. 726–736.

[14] Madhukar Pai et al. "Tuberculosis". en. In: *Nat Rev Dis Primers* 2 (Oct. 2016), p. 16076.

[15] Remco R Bouckaert and Alexei J Drummond. "bModelTest: Bayesian phylogenetic site model averaging and model comparison". In: *BMC evolutionary biology* 17 (2017), pp. 1–11.

[16] Xavier Didelot et al. "Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks". en. In: *Mol. Biol. Evol.* 34.4 (Apr. 2017), pp. 997–1007.

[17] Don Klinkenberg et al. "Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks". en. In: *PLoS Comput. Biol.* 13.5 (May 2017), e1005495.

[18] Finlay Campbell et al. "Outbreaker2: a modular platform for outbreak reconstruction". In: *Bmc Bioinformatics* 19 (2018), pp. 1–8.

[19] Finlay Campbell et al. "When are pathogen genome sequences informative of transmission events?" en. In: *PLoS Pathog.* 14.2 (Feb. 2018), e1006885.

[20] Marc A Suchard et al. "Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10". In: *Virus Evolution* 4.1 (2018), vey016.

[21] Remco Bouckaert et al. "BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis". In: *PLoS computational biology* 15.4 (2019), e1006650.

[22] Matthew D Hall and Caroline Colijn. "Transmission Trees on a Known Pathogen Phylogeny: Enumeration and Sampling". en. In: *Mol. Biol. Evol.* 36.6 (June 2019), pp. 1333–1343.

[23] Yuanwei Xu et al. "High-resolution mapping of tuberculosis transmission: Whole genome sequencing and phylogenetic modelling of a cohort from Valencia Region, Spain". en. In: *PLoS Med.* 16.10 (Oct. 2019), e1002961.

[24] Torsten Seemann et al. "Tracking the COVID-19 pandemic in Australia using genomics". en. In: *Nat. Commun.* 11.1 (Sept. 2020), p. 4376.

[25] Xavier Didelot et al. "Genomic epidemiology analysis of infectious disease outbreaks using TransPhylo". In: *Current protocols* 1.2 (2021), e60.

[26] Remco R Bouckaert. "An efficient coalescent epoch model for Bayesian phylogenetic inference". In: *Systematic Biology* 71.6 (2022), pp. 1549–1560.

[27] Helene Duault, Benoit Durand, and Laetitia Canini. "Methods combining genomic and epidemiological data in the reconstruction of transmission trees: a systematic review". In: *Pathogens* 11.2 (2022), p. 252.

[28] Martin R Smith. "Robust analysis of phylogenetic tree space". In: *Systematic Biology* 71.5 (2022), pp. 1255–1270.

[29] Benjamin Sobkowiak et al. "Comparing transmission reconstruction models with Mycobacterium tuberculosis whole genome sequence data". In: *bioRxiv* (2022), pp. 2022–01.

[30] Chongguang Yang et al. "Phylogeography and transmission of M. tuberculosis in Moldova: A prospective genomic analysis". In: *PLoS Medicine* 19.2 (2022), e1003933.

[31] Galo A Goig et al. "Effect of compensatory evolution in the emergence and transmission of rifampicin-resistant Mycobacterium tuberculosis in Cape Town, South Africa: a genomic epidemiology study". en. In: *Lancet Microbe* 4.7 (July 2023), e506–e515.

[32] Lars Berling et al. "A tractable tree distribution parameterized by clade probabilities and its application to Bayesian phylogenetic point estimation". In: *bioRxiv* (2024), pp. 2024–02.

[33] Jake Carson et al. "Inference of Infectious Disease Transmission through a Relaxed Bottleneck Using Multiple Genomes Per Host". en. In: *Mol. Biol. Evol.* 41.1 (Jan. 2024).

[34] Fabio K Mendes et al. "How to validate a Bayesian evolutionary model". In: *bioRxiv* (2024), pp. 2024–02.

A. Supplementary Materials

S1.1. Within-host coalescent likelihood

The second term in the decomposition in (4), $P(G|T, N_e g, \tau_s)$, is the likelihood of the phylogeny conditional on the transmission tree and the within-host coalescent parameter. As in previous work (BEASTLIER, TransPhylo, phybreak), computing this likelihood is aided by the fact that the transmission tree breaks the phylogeny into smaller, independent subtrees (one for each host), each occurring in a different individual, along with linear chains of transmission which are "trivial" genealogies of just one lineage. Accordingly, the phylogeny likelihood is the product of likelihoods of smaller phylogenies; we use a coalescent model for the within-host evolution.

Following [TP1], Let G_i be a timed genealogy corresponding to a single host, and let τ_i be the times of the leaves for this tree. (The sampling event for a sampled host's genealogy will be one of the leaf times, and also one of the times τ_s. Other tips in host i's genealogy correspond to times at which i infected another individual.) Let t_i^{inf} be the time of infection of host i. We model a strict bottleneck at transmission, and consequently, all lineages in G_i must coalescence after t_i^{inf} (forward in time), or before t_i^{inf} (thinking backward in time).

Still following [TP1], suppose there are n_i leaves in G_i, ordered from the earliest to the oldest. Consider a process by which we add each tip in the order in which it occurs. When the j'th tip arrives, it must coalesce with the genealogy formed by the first $j-1$ tips, and it must do so before time t_i^{inf}. Let A_j denote the sum of branch lengths in the genealogy formed, at the time that we add j, between the time of leaf j and the time where it coalesces with an ancestor of a previously considered leaf. Let B_j be the sum of branch lengths between the time of leaf j and t_i^{inf}. Then A_j is exponentially distributed, with parameter $N_e g$, but we must account for the fact that all coalescent events need to occur before t_i^{inf}, or in other words, $A_j < B_j$. This means that, if there are M non-trivial individual genealogies making up the genealogy G, we have

$$P(G|T, \theta, N_e g, \tau_s) = \prod_{i=1}^{M} P(G_i|N_e g, \tau_i) = \prod_{i=1}^{M} \prod_{j=1}^{n_i} \frac{\exp(-A_j/N_e g)}{N_e g(1-\exp(-B_j/N_e g))}.$$

A "trivial" genealogy consists of only one lineage (a line, with no branching events). Its probability is 1 in the coalescent model.

S1.2. Derivation of p_0

We obtain p_0, the probability of being unsampled and having all descendants unsampled (i.e. of not being ATTS) with a standard technique from branching processes: conditioning on the number of descendants. Because we will deal with right truncation momentarily, here, we let p_0 be the probability of *ever* being known:

$$p_0 = (1 - C^s) \sum_{k=0}^{\infty} p(k) p_0^k = (1 - C^s) g(p_0)$$

where $g(s)$ is the probability generating function for the offspring distribution $p(k)$. In our process, this is Poisson with mean C^{tr} if sampling does not stop infectors from infecting others after sampling (otherwise, it is not clear, but a negative binomial distribution would be a reasonable assumption, with a mean that is less than C^{tr}). For the Poisson distribution with mean λ, $g(s) = e^{\lambda(s-1)}$. We use Newton's method to compute p_0 numerically.

Now we can build the conditioning term $1 - S^E(Y_R)$.

S1.3. Derivation of ϕ

The probability that an individual is not ancestral to the sample (ATTS) is p_0. The probability that an indvidual j is "once ATTS" is the probability that precisely one of j's infectees is ATTS, or

$$\begin{aligned}
P(1 \text{ ATTS}) &= \sum_{k=1}^{\infty} p(k)k(1-p_0)p_0^{k-1} \\
&= \sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!}k(1-p_0)p_0^{k-1} \\
&= \lambda(1-p_0)\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}p_0^{k-1} \\
&= \lambda(1-p_0)\sum_{m=0}^{\infty}\frac{\lambda^m}{m!}p_0^m \\
&= \lambda(1-p_0)g(p_0) \\
&= \lambda(1-p_0)\frac{p_0}{1-C^s}
\end{aligned} \tag{11}$$

where g is the probability generating function for the distribution $p(k)$, which we model as Poisson, and the final line follows from the form of p_0 in the main text. In order to be what we have called "twice-ATTS", an individual must simply not be unknown, and not be only once ATTS, so we have

$$\phi = 1 - p_0 - \lambda(1-p_0)\frac{p_0}{1-C^s}$$

which is what is in the main text.

S1.4. Details of the block likelihood

Let $h^e(t) = h^s + h^{tr}\phi$, where ϕ is the probability that the individual is MATTS. This probability is $\phi = 1 - p(0 \text{ ATTS}) - p(1 \text{ ATTS})$:

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

After some manipulation (see Appendix S1.3), we have

$$\phi = 1 - p_0\left(1 + \frac{\lambda(1-p_0)}{1-C^s}\right)$$

where λ is the mean of the offspring distribution $p(k)$. This is C^{tr} if sampling does not prevent onward transmission, and something less than C^{tr} if it does. We leave the question of adapting for sampling's effect on transmission for later.

Now we can build the success probability for our geometric distribution. The probability ρ that an event from the intensity function h^e happens is one minus the probability that it never occurs:

$$\rho = 1 - \exp\left(-\int_0^{\infty}(\phi h^{tr}(s) + h^s(s))ds\right) = 1 - S^{tr}(\infty)^{\phi}S^s(\infty)$$

We need ∞ in the argument, rather than a right censoring time, because we need to account for right truncation using the general principle $f(t|t < Y_R) = f(t)/(1-S)Y_R)$, as described in the main text, so we need $f(t)$, the density for the time if there were no truncation. Accordingly, in the numerator we need the density without adjustment for the finite time, and in the conditioning, we need to account for Y_R, which in this host is $\tau_{\text{end}} - \tau_i^k$ (the start time of the block).

Here, $p(n,t) = p(n)p(t|n)$ where $p(n) \sim \text{geom}(\rho)$ and $t = \sum_{i=1}^n t_i$, where t_i is the length of time it takes for an individual in the chain to infect the next individual. We model $t_i \sim \text{gamma}(a, b)$, so that $t \sim \text{gamma}(na, b)$.

The likelihood is:

$$L(n, t|\theta) = \frac{p(n,t)}{1 - P(t > Y_R^j)} \tag{12}$$

where n is the count, Y_R^j is the block's right truncation time and t is the duration of the block.

The term in the denominator, $P(t \leq Y_R^j)$ is

$$P(t \leq Y_R^j) = \sum_n p(n)p(t \leq Y_R^j|n).$$

Using a geometric distribution for $p(n)$ and $\gamma(na, b)$ for t, we have

$$P(t \leq Y_R^j) = \sum_{n=1}^{\infty} (1-\rho)^n \frac{\gamma(na, bY_R^j)}{\Gamma(na)}$$

where $\gamma(s, x)$ is the incomplete lower gamma function, $\gamma(s,x) = \int_0^x t^{s-1}e^{-t}dt$, and $\Gamma(x)$ is the Gamma function, $\Gamma(x) = \int_0^\infty t^{z-1}e^{-t}dt$. This comes from the density for the gamma-distributed random variable x with shape parameter a and rate b, namely $f(x) = \frac{b^a}{\Gamma(a)}x^{a-1}\exp(-bx)$. The CDF for the gamma distribution with shape na and rate b (evaluated at Y_R^j) gives us $\frac{\gamma(na, bY_R^j)}{\Gamma(na)}$ directly. We compute the sum numerically with a tolerance of 10^{-7} and capped n at a million. We have now defined all the ingredients needed to explicitly write, and compute, the transmission likelihood in (5), and in fact, the whole likelihood in (4). We use $a = A^{tr}$ and $b = B^{tr}$.

S1.5. Simulator

We also implemented a simulator independent of the transmission likelihood, that allows us to simulate transmission trees with phylogenies, allowing us to test the model. The inputs for the simulator are the parameters for the sampling and transmission intensity functions (3 each), $N_e g$ and the stopping time t_e.

Initialize a list L of hosts whose infections are to be simulated. L can contain the times of infection, and the infector. Start with one host to be simulated, starting at time $t = 0$. L starts out being a row $(0, \text{NA})$. (NA for the fact that this is the index host, and they have no infector in the process).

Simulate an individual host, i, with time of infection t_i:

1. Simulate $n_i \sim Poisson(C^{tr})$. This is the number of new infections that the host would eventually cause (without considering the stopping time).
2. Simulate whether the host will be sampled: $s_i \sim \text{Bernoulli}(p)$, where $p = C^s$. If $s_i = 1$ the host would be sampled, if the sampling time occurs early enough.
3. If $s_i = 1$, simulate the time of sampling:

$$t_i^s - t_i \sim \text{gamma}(A^s, B^s)$$

If $t_i^s > t_e$, the host is not sampled after all. Ignore t_i^s.

4. Simulate the times when i infects the n_i new infectees:

$$t_i^k - t_i \sim \text{gamma}(A^{tr}, B^{tr})$$

with $k = 1, ..., n_i$. Remove any t_i^k that are above t_i^s (model A, where sampling stops people from transmitting) or t_e (model B, where transmission ends when we end observation).

5. Sampling a within-host phylogeny (see below) for host i using the times $t_i^{tips} = \{t_i^k, \text{ (if sampled) } t_i^s\}$. Attach the origin of this phylogeny to the infector of host i, accounting for the fact that the tMRCA of this phylogeny is *after* (forwards in time) t_i, not equal to t_i.

6. Remove the (t_i, infector for t_i) row from L. Append the set $\{t_i^k : t_i^k < t_e\}$ times of infection and i as the infector to the list L of infection times and infectors. L gets rows (t_i^k, i) appended to it, only for those t_i^k that are less than t_e.

Continue by simulating the hosts in the list, gluing the phylogenies together. Stop when there are no more hosts to infect. Depending on the intensity function parameters, trees with zero or one taxon often have a high probability, which can be interpreted as an outbreak immediately being stopped. Another mode can often be observed with just a few taxa, indicating the outbreak dies out very early on. But there tends to be a long tail in the taxon count distribution, some reaching a very high number of taxa. If a fixed number of taxa is desired, the tree is rejected if the number of taxa differs from the desired one, and a new tree is generated until one is encountered with the desired number of taxa. Due to the thinness of the tail and depending on the intensity functions' parameters, this can take considerable time. This also has the potential to produce bias in some tests of the model, because the desired number of taxa may be unlikely given the epidemiological parameters. Our simulation tests do not seem affected by the choice of the number of taxa.

Some of the within-host phylogenies will be trivial. These are unsampled hosts who infect precisely one other, who is not sampled. These become the chains of unsampled transmission. Any host (sampled or not) with two or more tips in t_i^{tips} contains at least one node of the larger phylogeny, because it has at least one coalescent event in it.

Figure S6 shows a simulated tree including within host coalescent events and unsampled hosts. The red tree is the one output by the simulator.

The transmission tree simulator is available as the `TransmissionTreeSimulator` app in the `BREATH` package for BEAST 2. It has the following options:

- endTime (real number): end time of the study
- popSize (real number): population size governing the coalescent process
- sampleShape (real number): shape parameter of the sampling intensity function
- sampleRate (real number): rate parameter of the sampling intensity function
- sampleConstant (real number): constant multiplier of the sampling intensity function
- transmissionShape (real number): shape parameter of the transmission intensity function
- transmissionRate (real number): rate parameter of the transmission intensity function
- transmissionConstant (real number): constant multiplier of the transmission intensity function
- out (file name): output file. Print to stdout if not specified (optional)
- trace (file name): trace output file, or stdout if not specified (optional)

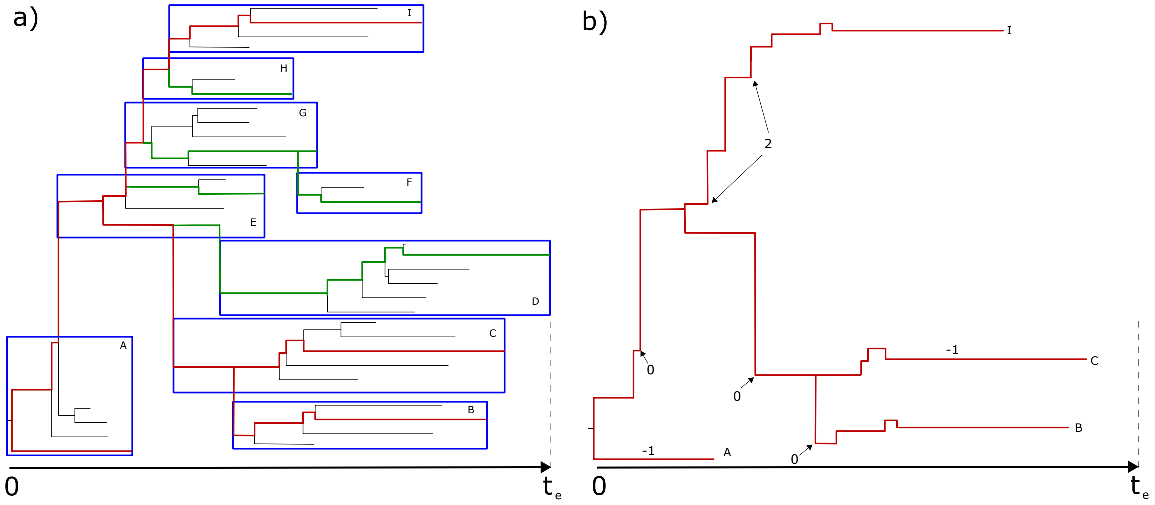


Figure S6: Simulated tree from time 0 to time t_e indicated on the x-axis. a) Small simulated tree where blue boxes indicate hosts, red tree the tree ending in samples, red+green branches are branches generated by the simulator as within host coalescent trees, red+green+black branches form the underlying phylogeny. b) Tree output by the simulator. Numbers on branches are block counts, arrows indicate block start and end of blocks (if block count at least zero). For blocks with count of zero start and end of block coincide. Hosts D to H are not sampled, so these are removed from the simulator output. Host E becomes an unsampled host infected by A and infecting C. Hosts G and H form a block of size 2 of unknown hosts, while hosts D and F leave no trace and remain unknown unknowns.

- seed (long): random number seed used to initialise the random number generator (optional)
- maxAttempts (integer): maximum number of attempts to generate coalescent sub-trees (default: 1000)
- taxonCount (integer): generate tree with taxonCount number of taxa. Ignored if negative (default: -1)
- maxTaxonCount (integer): reject any tree with more than this number of taxa. Ignored if negative (default: -1)
- treeCount (integer): generate treeCount number of trees (default: 1)
- directOnly (true|false): consider direct infections only, if false block counts are ignored (default: true)
- quiet (true|false): suppress some screen output (default: false)

To use the command line version of the simulator, use the ‘applauncher‘ application (which is part of the BEAST 2 distribution) from a terminal/command prompt. Any of the above options can be used.

Alternatively, start BEAUti (which is also part of the BEAST 2 distribution), select the ‘File/Launch apps‘ menu, and select ‘TransmissionTreeSimulator‘ from the list of applications. Click the ‘launch‘ button to start a GUI version of the simulator.

There is a `WIWVisualiser` tool available in the `BREATH` package to create SVG files to visualise who infected whom, based on a posterior sample. The `WIWVisualiser` has the following options:

- trees (file name): tree file file with transmission trees (optional).
- log (file name): trace file containing infectorOf log. Ignored if tree file is specified (optional)
- burnin (integer): percentage of trees to used as burn-in (and will be ignored) (optional, default: 10)
- out (file name): output file, or stdout if not specified (optional, default: /tmp/wiw.svg)
- prefix (string): prefix of infectorOf entry, e.g., infectorOf (optional, default: infectorOf)
- threshold (real number): probability threshold below which edges will be ignored. (optional, default: 0.1)
- partition (string): name of the partition appended to 'blockcount, blockend and blockstart' (optional)

S2. Comparison to TransPhylo

We ran TransPhylo using the same input phylogeny as was used in the original TransPhylo paper [16]. (We note that TransPhylo has a bug in which if the date of last possible sampling is set to the last tip date, the model may place too many unsampled cases near the tips of the tree). We used the following command, i.e. 100,000 iterations, updating the sampling probability, offspring distribution parameter r but not p and with a date of last possible sampling set to 2015:

```
roetztree=read.tree("roetz.nwk")
neg=0.03;

recnew=inferTTree(ptreeFromPhylo(roetztree,dateLastSample = 2010.91),w.shape=1.7,
                  w.scale=1/0.3, ws.shape=1, ws.scale=0.5,
                  mcmcIterations=100000,thinning=30, startNeg=neg,startPi=0.5,
                  startOff.r=1,updateOff.r=T,updatePi=T,updateNeg=T,updateOff.p=F,
                  dateT =2015, updateTTree = T,optiStart=T)
```

We checked parameter traces for convergence, and extracted who infected whom information with the 'computeMatWIW' function with a 20% burnin. Likely because of the fixed phylogeny assumption, TransPhylo produces more high-probability pairs among sampled cases than BREATH. TransPhylo's fixed phylogeny places constraints on the transmission tree, which of course are not constraints in BREATH which can change the phylogeny. Figure S7 shows the empirical cumulative distribution functions. Both methods, naturally, have a large number of very small probability events, which is due to the constraint of a transmission tree: there are $\binom{N}{2}$ possible pairs and each individual has at most one infector among the sampled individuals, so there are at most $N-1$ transmission events among sampled individuals in each posterior tree.

Figure S9 compares the posterior probabilities for the highest-probability sources for each sampled host in BREATH with the posterior probability for that same soutce in TransPhylo. The highest probability transmission pairs in BREATH (there are 7 with probability above 0.5) also have high probability in TransPhylo, but the converse is not the case. Figure S8 shows the posterior probabilities for sampled pairs and the genetic distances between the pairs. Both methods have higher overall probabilities for pairs whose TB sequences are more similar (low distance), but BREATH shows a more clear pattern of increasing probability as the SNP distance goes down. This is also likely due to the ability to update the phylogenetic tree. Figure S10 shows the

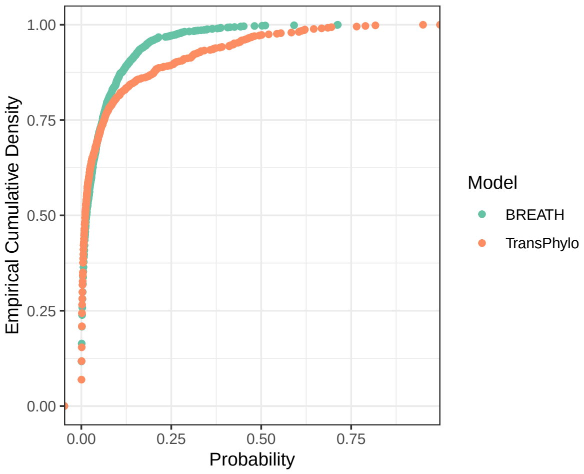


Figure S7: Empirical cumulative distribution functions for the posterior probabilities that i infected j for all sampled pairs of hosts i, j (with $i \neq j$).

posterior probabilities of each possible source for each recipient, including an unsampled source, according to both models. The BREATH posterior is notably more diffuse, again demonstrating the artificial certainty imposed by a fixed phylogenetic tree.

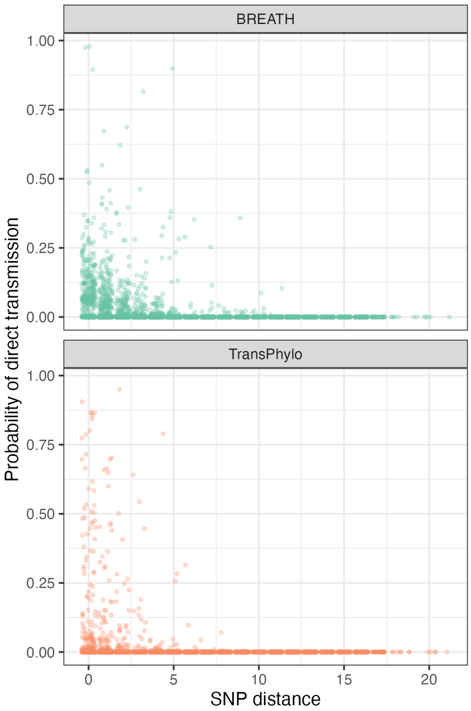


Figure S8: Posterior probabilities for pairs of sampled individuals (the sum of both source-recipient directions for the pair) vs single-nucleotide polymorphism (SNP) distances between the individuals' TB genomes in BREATH (top) and TransPhylo (bottom).

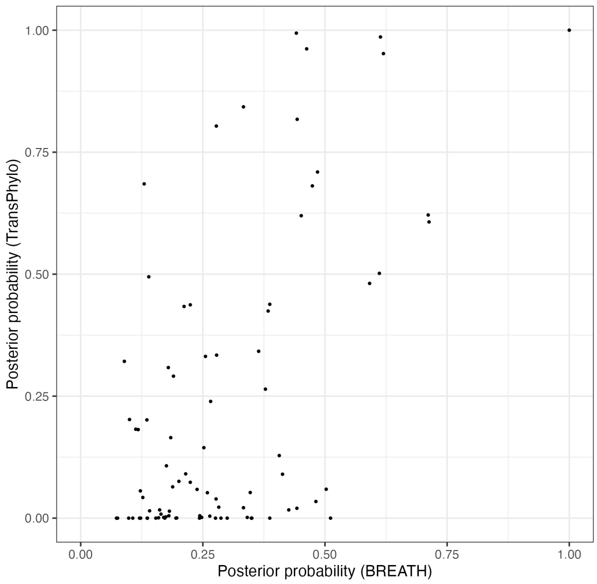


Figure S9: Scatter plot comparing posterior probabilities for the highest-probability source for each source in BREATH (x-axis) to the probability of the same source from TransPhylo, regardless of whether it was the highest-probability source in TransPhylo. Unsampled sources are included.

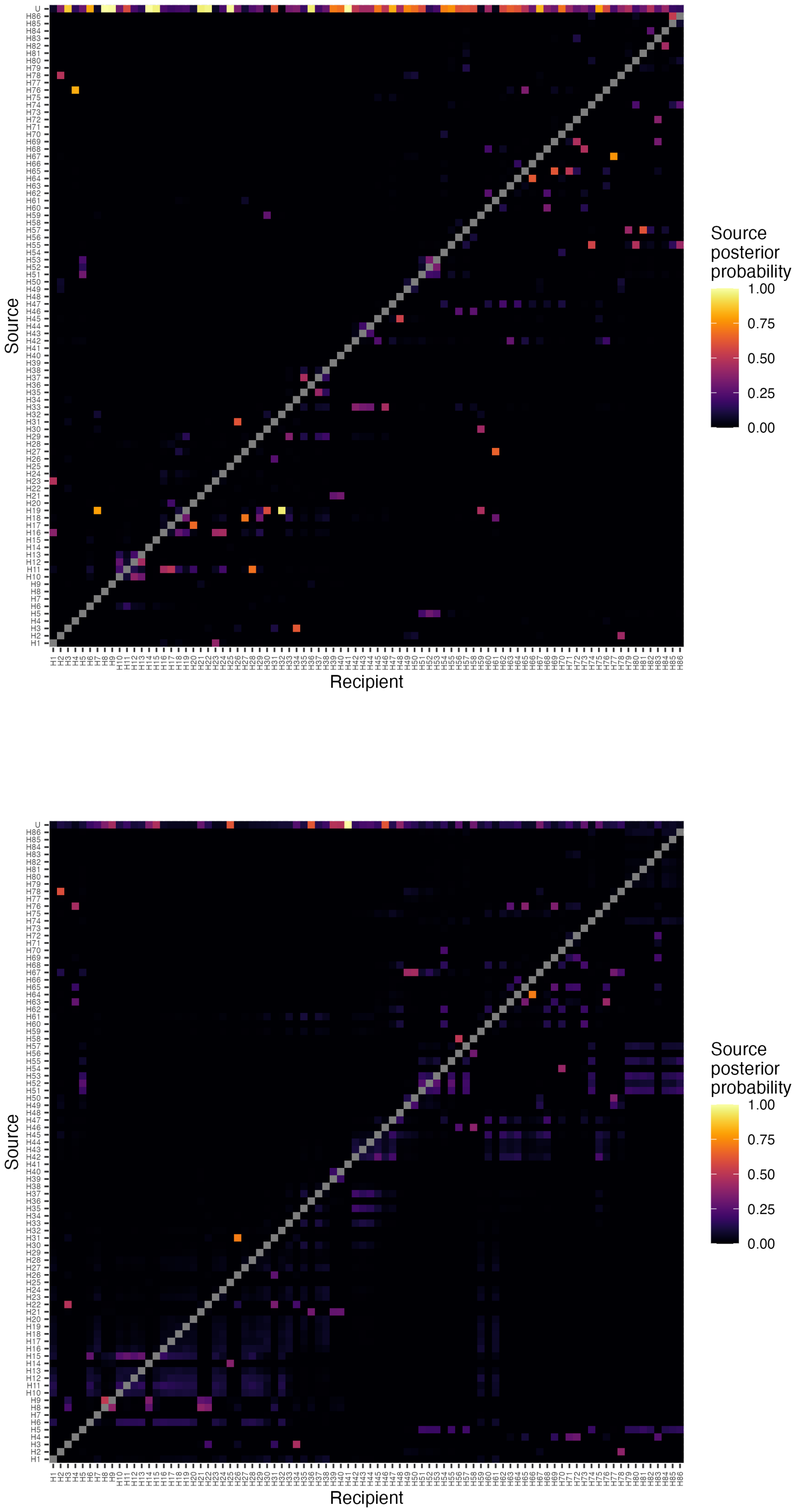


Figure S10: Heatmaps for posterior probabilities of the infector (y-axis) of all sampled cases (x-axis). "U" represents infection by an unsampled host. Top: TransPhylo. Bottom: BREATH.