*Statistical Science*
2018, Vol. 33, No. 1, 70–85
https://doi.org/10.1214/17-STS637

# Estimating Transmission from Genetic and Epidemiological Data: A Metric to Compare Transmission Trees

**Michelle Kendall, Diepreye Ayabina, Yuanwei Xu, James Stimson and Caroline Colijn**

*Abstract.* Reconstructing who infected whom is a central challenge in analysing epidemiological data. Recently, advances in sequencing technology have led to increasing interest in Bayesian approaches to inferring who infected whom using genetic data from pathogens. The logic behind such approaches is that isolates that are nearly genetically identical are more likely to have been recently transmitted than those that are very different. A number of methods have been developed to perform this inference. However, testing their convergence, examining posterior sets of transmission trees and comparing methods' performance are challenged by the fact that the object of inference—the transmission tree—is a complicated discrete structure. We introduce a metric on transmission trees to quantify distances between them. The metric can accommodate trees with unsampled individuals, and highlights differences in the source case and in the number of infections per infector. We illustrate its performance on simple simulated scenarios and on posterior transmission trees from a TB outbreak. We find that the metric reveals where the posterior is sensitive to the priors, and where collections of trees are composed of distinct clusters. We use the metric to define median trees summarising these clusters. Quantitative tools to compare transmission trees to each other will be required for assessing MCMC convergence, exploring posterior trees and benchmarking diverse methods as this field continues to mature.

*Key words and phrases:* Infectious diseases, genomics, epidemiology, Bayesian inference, modelling.

## 1. INTRODUCTION

Understanding who infected whom is a key task of epidemiology. High quality reconstruction of who infected whom in an outbreak of an infectious disease allows public health workers to determine whether there are individuals or locations causing high numbers of transmission, to identify those individuals at risk, and to determine which individual characteristics are associated with infectiousness. Ultimately, this knowledge leads to improved infection control and outbreak man-

*Michelle Kendall is Research Associate, Department of Mathematics, Huxley Building, 180 Queen's Gate, South Kensington Campus, London, SW7 2AZ, United Kingdom (e-mail: m.kendall@imperial.ac.uk). Diepreye Ayabina is Research Student, Department of Mathematics, Huxley Building, 180 Queen's Gate, South Kensington Campus, London, SW7 2AZ, United Kingdom (e-mail: d.ayabina14@imperial.ac.uk). Yuanwei Xu is Research Associate, Department of Mathematics, Huxley Building, 180 Queen's Gate, South Kensington Campus, London, SW7 2AZ, United Kingdom (e-mail: yuanwei.xu@imperial.ac.uk). James Stimson is Research Associate, Department of Mathematics, Huxley Building, 180 Queen's Gate, South Kensington Campus, London, SW7 2AZ, United Kingdom (e-mail: james.stimson16@imperial.ac.uk). Caroline Colijn is Reader, Department of Mathematics, Huxley Building, 180 Queen's Gate, South Kensington Campus, London, SW7 2AZ, United Kingdom (e-mail: c.colijn@imperial.ac.uk).*

agement. However, outbreak reconstruction is time-consuming, expensive and uncertain. It often must rely on individuals' recollections of those with whom they have had contact, as well as individual health records, locations in which infection may have spread, and so on. Particularly in the case of sexually transmitted infections and blood-borne infection, this information is sensitive and case identification is challenging. For chronic infections, transmission may have occurred a considerable time before diagnosis, making reconstructing transmission even more challenging.

For these reasons and others, there is considerable interest in using genetic data from rapidly evolving viruses and even bacteria in outbreak reconstructions. Recent advances in sequencing technology have meant that it is feasible to obtain whole-genome RNA or DNA sequences from pathogens even in real time during outbreaks [30, 11], and these data can be used to perform outbreak reconstructions, or to refine reconstructions based on traditional epidemiology. The central idea behind genomic approaches to outbreak reconstruction is that genetic polymorphisms in viruses or bacteria accrue even in the short time frame of the outbreak; by comparing cases' pathogen sequences, it is possible to refine estimates of who infected whom. For example, if cases A and B were in close contact at a time when A was infectious, epidemiological investigations alone would likely conclude that A infected B, but if the pathogen sequences are very different genetically, it would rule this out and another infector would be sought to explain B's infection.

However, inference of transmission using genetic sequences is challenging. It relies not only on a knowledge of the likely time between an individual becoming infected and infecting others (the generation time), and on the likely time between becoming infected and seeking treatment (leading to being known to the health care system)—this information is used in almost any reconstruction of transmission. Incorporating genetic data also requires a model of how mutations occur: at the time of transmission, or continuously throughout the life of the pathogen, and at what rate (clocklike evolution or a more general model). It requires, implicitly or explicitly, a model of the dynamics of the pathogen within and between hosts: is more than one lineage present, and how many pathogen particles are transmitted upon infection? Finally, it is rare that health authorities identify every case in an outbreak, and handling unknown cases raises additional challenges. Ideally, genetic information is integrated with epidemiological and clinical information to obtain the best possible estimates of who infected whom.

Interest in the statistical tools necessary to solve these problems is growing rapidly, and diverse methods have been developed. These differ in their statistical approach: whether they have an explicit spatial structure [27, 26]; whether they allow multiple introductions of the pathogen into the community being analysed [18, 26, 38], or not [28, 36]; whether they do not allow multiple distinct infections of individual hosts [38, 15]; whether they consider the population dynamics of the pathogen in the host [23, 36], or not [18, 27]; whether they use a phylogenetic tree to capture relationships among the pathogen sequences [8, 9, 21], or infer the phylogenetic tree and transmission tree simultaneously [15, 23, 7]; and whether they handle the issue of unknown cases and/or cases without genetic data [8, 23, 18, 15]. Table 1 lists some of the available tools with respect to these variations. While there are a number of exemplars illustrating the relationship between genomic data and transmission (examples include [37, 14, 24, 12]), we focus on Bayesian inference methods aiming to provide tools for use by the community.

The data integration needs of this field motivate the use of Bayesian approaches, as they provide a natural framework for integration of covariates such as location, clinical indications of infectiousness and other variables, and avoid the need to use summary statistics of the data. However, by their nature, Bayesian approaches produce a posterior collection of inferred transmission trees alongside posterior distributions of scalar parameters. Understanding the nature of posterior uncertainty in a complex object such as a transmission tree is not straightforward. For example, do posterior estimates group into some trees in which case A was infected first (we say "A is the source"), A then infected B, and B went on to infect several others, ultimately causing the outbreak, versus trees in which case B is the source, B infected D, and D then caused the other infections? Do the data support distinct alternative stories of the outbreak, or is the posterior unimodal in the space of transmission trees? Which transmission chains had more unsampled cases? Typically, the fraction of correctly inferred infectors, or the fraction consistent with an external set of data, is used as a measure of the quality of inferred transmission trees. However, this does not capture "how wrong" the incorrect links are, and does not allow informative comparisons either within a posterior set of trees or of the performance of different methods. In addition, summarising the posterior is typically achieved using the Edmond's consensus tree [13, 9, 23]: a consensus graph is constructed by finding the most common infector for each infectee,

TABLE 1

*Some available methods for reconstructing transmission trees using genetic data. "Multi. intro" refers to whether the method accounts for multiple introductions of a pathogen into a community, distinguishing whether all cases are part of one outbreak or several smaller ones. "Multi. seq" refers to whether the method allows for more than one sequenced isolate per case; often this does not mean multiple distinct infections (re-infection), but only monophyletic clonal instances. "In-host" refers to whether the method admits pathogen diversity within individual hosts; if yes, coalescent or branching events may not correspond to transmission events. "Unsamp" refers to whether there may be inferred cases that were not known to health authorities and not included in the dataset (in contrast to known cases without sequences). "Bneck > 1" refers to whether pathogen diversity can be transmitted (if yes) or whether only one unique sequence is transmitted from case to case (if no). "Phy. Tree" refers to whether a phylogenetic tree is required as an input (if Yes), estimated alongside transmission (if Est.), or not used (if No). "Seqs" refers to whether genetic sequences are used directly in the inference procedure. "Exp. Time" refers to whether data concerning time of exposure to disease or length of admission time is used in the inference procedure. "Loca. data" refers to whether location data is used in the inference procedure*

| | | **Method features** | | | | | **Data Used** | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Name/Author** | **Ref** | **Multi intro** | **Multi Seq.** | **In host** | **Bneck > 1** | **Un Samp** | **Phy Tree** | **Seqs** | **Exp Tim** | **Loca. data** |
| Outbreaker | [18] | Yes | No | No | No | Yes | No | Yes | No | No |
| TransPhylo | [8] | No | Yes | Yes | No | Yes | Yes | No | No | No |
| SCOTTI | [7] | Yes | Yes | Yes | Yes | Yes | Est. | Yes | Yes | No |
| Kenah et al. | [21] | No | Yes | Yes | No | No | Yes | No | Yes | No |
| Numinnen et al. | [28] | No | Yes | Lim. | No | No | Est. | Yes | No | No |
| Mollentze et al. | [26] | Yes | No | No | No | No | Yes | No | Yes | Yes |
| Morelli et al. | [27] | No | No | No | No | Yes | No | Yes | Yes | Yes |
| Soubeyrand | [35] | No | No | Yes | Yes | No | Yes | No | Yes | Yes |
| Hall et al. | [15] | No | Yes | Yes | No | Yes | Est. | Yes | No | Yes |
| phybreak | [23] | No | Yes | Yes | No | No | Est. | Yes | No | No |
| Trepar | [36] | No | No | Yes | No | Yes | Yes | No | No | No |
| bitrugs | [38] | Yes | No | Yes | No | Yes | No | Yes | Yes | No |

and then Edmond's algorithm is used to find the minimum directed spanning tree of this graph. It is therefore possible that such a consensus tree is different in structure from every tree in the posterior, particularly when the trees are quite varied. This limits the ability to effectively summarise the posterior.

Here, we develop a metric on the space of transmission trees for a set of infected cases. It allows for unsampled individuals in transmission trees, and is also applicable to other kinds of tree structures. We illustrate the metric using random transmission trees with a simple structure, and find that the metric separates groups of transmission trees in an intuitive and meaningful way. We proceed to analyse posterior collections of transmission trees from a Bayesian inference of transmission from genetic data, and we illustrate how the metric allows us to understand posterior uncertainty and sensitivity to priors. Additionally, the metric provides a straightforward way to identify a representative median tree from a collection of trees. Such a median tree has advantages over consensus tree constructions because it is one of the trees from the original collection.

## 2. THE METRIC

We begin by defining what we mean by a transmission tree. We consider the case in which each individual is infected at most once. For many pathogens, it is possible that cases are infected sequentially or even coinfected with different variants, but if this is observed in a set of data, we would denote the multiple infections as distinct, each with a unique infector. Note that we allow for the presence of *unsampled* cases among the nodes, that is, individuals who were not known to the health care system during the data-gathering process, but whose presence in the transmission has been inferred.

DEFINITION 1. A *transmission tree* $T = (N, E)$ is a directed graph with nodes $N$ and edges $E$, in which each node corresponds to an infected individual and edges correspond to transmission events. The set of nodes $N = S \cup U$, where $S$ is the set of sampled cases and $U$ is the (possibly empty) set of unsampled cases. A directed edge from node $n_i$ to $n_j$ implies that $n_i$ infected $n_j$. We say that $n_i$ is the "infector" and $n_j$ is the "infectee". Each node has at most one infector. We require the graph to comprise a single connected component. In addition, a transmission tree has a unique node, the *source*, with in-degree 0 (no infector in $N$).

Since we do not allow for an infectee to have more than one infector, and we have a unique source with no infector in $N$, and the graph is connected, the graph $(N, E)$ has no cycles; it is a tree.

DEFINITION 2. For any node $n_i \in N$, there is a unique path $p_i$ in $T$ from the source case along directed edges to $n_i$.

The *depth* of node $n_i$ is the number of edges on the path $p_i$; the source case has depth zero.

The *most recent common infector* (MRCI) of two nodes $n_i$ and $n_j$ is the node with the greatest depth which lies on both paths $p_i$ and $p_j$. Note that if $n_i$ infected $n_j$, or more generally if $n_i$ lies on the path $p_j$, then their MRCI is $n_i$. For convenience, the MRCI of $n_i$ and $n_i$ is also defined to be $n_i$.

The *descendants* of $n_i$ are the nodes that can be reached following directed paths originating at $n_i$.

The requirement that there is a unique source node reflects the fact that we are not modelling multiple distinct introductions of a pathogen into a community. Rather, the source node is infected somewhere, by someone, outside the study population and introduces the infection into the study population via the transmission tree.

DEFINITION 3. For a transmission tree $T$, we define the matrix $v(T)$ with components $v_{i,j}$ = the depth of the MRCI of $n_i$ and $n_j$ in $T$.

We illustrate a simple transmission tree and give some examples of $v$ in Figure 1.

To compare different transmission trees $T_1$ and $T_2$ for the same infection, we propose using the Euclidean distance between $v(T_1)$ and $v(T_2)$ (each written for convenience as a vector), as was done in [4, 22]. However, although the trees will contain the same set of sampled cases, $S = \{s_1, s_2, \dots, s_{|S|}\}$, the number of inferred unsampled cases $|U|$, and hence $|N|$, may differ between trees. Therefore, to ensure that we are comparing vectors of the same length, we restrict our attention to the vector of sampled cases,

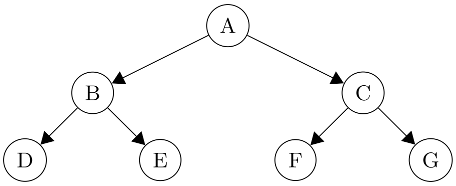


FIG. 1. *A simple transmission tree. Here,* $v_{D,E} = 1$ *because the MRCI of cases D and E is case B, which is* 1 *step from the source case A.* $v_{D,B} = 1$ *also, because the MRCA of D and B is B. But* $v_{D,F} = v_{E,F} = v_{E,G} = 0$, *and so on for pairs of cases whose MRCI is the source case, A.*

$$v|_S(T) = (v_{s_1,s_1}, v_{s_1,s_2}, \dots, v_{s_{|S|},s_{|S|}}).$$

In practice, we will often wish to compare trees with respect to transmission paths leading to sampled cases, ignoring sets of "trailing" unsampled cases with no sampled descendants. Indeed, many tree inference methods only include unsampled cases to make sense of historic infectors of sampled cases. The tree vector of sampled cases respects this.

LEMMA 1. *Let* $T = (N, E)$ *be a transmission tree. Let* $T^* = (N^*, E^*)$ *be a copy of* $T$, *except that any unsampled cases in* $T$ *without infectees have been pruned* (*i.e., the unsampled case node and its only incident edge removed*), *and this process repeated until each unsampled case has at least one sampled case somewhere among its descendants. Then* $v|_S(T) = v|_S(T^*)$.

PROOF. The vector $v|_S$ records the depths of sampled cases (the $v_{s_i,s_i}$ entries, where $s_i \in S$) and the depths of MRCIs of pairs of sampled cases. Recall that by "depth" we mean the number of edges (equivalently, the number of nodes minus one) on the unique path from the source case to the node in question. Consider an unsampled case $u$ with no sampled case descendants. Since its removal would not shorten any path between the source case and a sampled node, or even between the source case and the MRCI of any pair of sampled nodes, its existence and position are entirely masked from $v|_S$. Thus each entry of $v|_S(T)$ is unchanged by the pruning of unsampled cases without sampled descendants, and so $v|_S(T) = v|_S(T^*)$. □

Since we are interested in comparing transmission trees, it is important to establish when we consider two trees to be equivalent. In particular, since the labels of the sampled cases are key to understanding the transmission process, it is important to distinguish between a tree where "case 1 infected case 2" and a tree where "case 2 infected case 1". However, since any numbering of unsampled cases is arbitrary, the labels of unsampled cases may be safely ignored. We will use the following definition.

DEFINITION 4. Consider two transmission trees $T_1 = (N_1 = S \cup U_1, E_1)$ and $T_2 = (N_2 = S \cup U_2, E_2)$, where the set of sampled cases $S$ is the same in each tree. Let $T_1^* = (N_1^* = S \cup U_1^*, E_1^*)$ and $T_2^* = (N_2^* =$

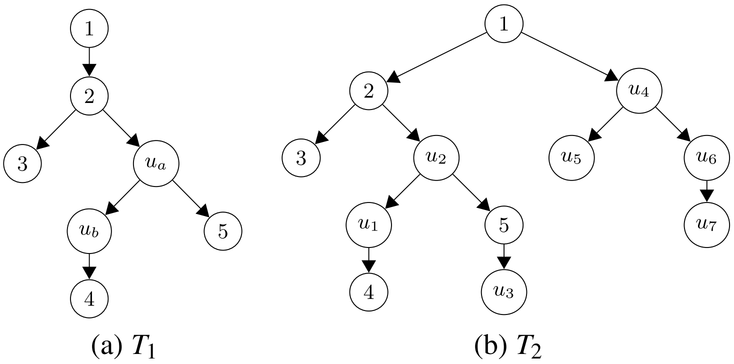


FIG. 2. *$T_1$ and $T_2$ are $S$-isomorphic because $T_2$ will be the same as $T_1$ (up to the relabelling of unsampled cases) after pruning the unsampled cases with no sampled descendants. Explicitly, we are using the bijection $\phi: N_1^* \to N_2^*$ where $\phi(s_i) = s_i$ for $s_i \in S = \{1,2,3,4,5\}$ and $\phi(u_a) = u_2$, $\phi(u_b) = u_1$.*

$S \cup U_2^*, E_2^*)$ be copies of $T_1$ and $T_2$, respectively, but pruned so that every unsampled case has at least one sampled case among its descendants, as in Lemma 1.

We say that $T_1$ and $T_2$ are *$S$-isomorphic* if there is an $S$-label-preserving isomorphism from $T_1^*$ to $T_2^*$, that is, a bijective function $\phi: N_1^* \to N_2^*$ such that $\phi$ is the identity on $S$:

$$\phi(s_i) = s_i \quad \text{for all } s_i \in S$$

and unpruned edges are preserved:

$$(n_i, n_j) \in E_1^* \quad \Leftrightarrow \quad \big(\phi(n_i), \phi(n_j)\big) \in E_2^*.$$

As an example, the two trees in Figure 2 are $S$-isomorphic: arbitrary differences in labelling of unsampled cases $u_1, u_2, \dots$ will not affect our measure of tree difference, nor will the presence of unsampled cases with no sampled descendants.

THEOREM 1. *Let $S$ be a set of sampled cases and $\mathcal{T}$ a set of transmission trees, each of whose set of nodes contains the set $S$. Then for any $T_1, T_2 \in \mathcal{T}$, the Euclidean distance between tree vectors,*

$$d(T_1, T_2) = \|v|_S(T_1) - v|_S(T_2)\|,$$

*is a metric on $\mathcal{T}$ up to $S$-isomorphism.*

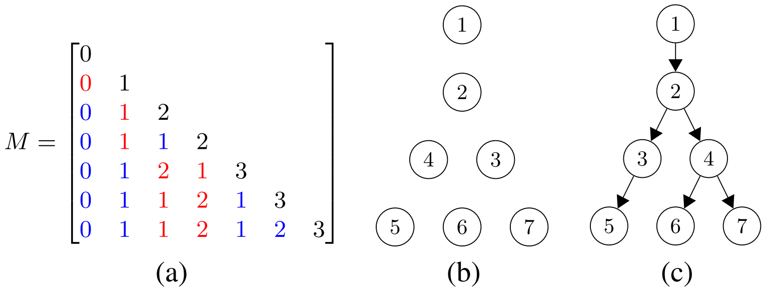
PROOF. The Euclidean distance between vectors is symmetric, nonnegative and satisfies the triangle inequality. To prove that $d$ is a metric, we need to show that $d(T_1, T_2) = 0$ if and only if $T_1$ and $T_2$ are $S$-isomorphic.

Since the vectors are well defined and are not conditional on the labelling of unsampled cases, and by Lemma 1, we know that when $T_1$ and $T_2$ are $S$-isomorphic then $v|_S(T_1) = v|_S(T_2)$. It remains to show that $v|_S(T_1) = v|_S(T_2)$ implies that $T_1$ and $T_2$ are $S$-isomorphic. The proof follows fairly naturally from results in [4] and [22]. Here, we provide a proof which also supplies some intuition for an algorithm for reconstructing the transmission tree $T$ from the tree vector $v|_S(T)$.

Let $T_1 = (N_1, E_1), T_2 = (N_2, E_2) \in \mathcal{T}$ be trees on a set of sampled cases $S$, and suppose that $v|_S(T_1) = v|_S(T_2)$. First, we consider the simpler case where there are no unsampled nodes in either tree, so $N_1 = N_2 = S$. We consider the identity bijection $\phi: N_1 \to N_2$ with $\phi(n_i) = n_i$ for all $i \in S = N_1 = N_2$. To show that $T_1$ and $T_2$ are $S$-isomorphic, we must show that $\phi$ preserves all edges so that $E_1 = E_2$.

The unique node $n_0$ with $v_{0,0}(T_1) = 0$ is the source case in $T_1$, and for each $i \in N_1$, the value $v_{i,i}(T_1)$ gives the depth of node $n_i$ in $T_1$; similarly for $T_2$. Thus $v|_S(T_1) = v|_S(T_2)$ implies that $T_1$ and $T_2$ have the same source case and that each sampled node is found at the same depth in both trees. We begin to see how the vector $v|_S(T_1)$ can be used to construct $T_1$: for each depth $\delta$, we can make a list of the nodes at that depth [nodes $n_j$ which satisfy $v_{j,j}(T_1) = \delta$]. In this way, we can start to draw our transmission tree as in Figure 3(b), where nodes are at the correct depths but directed edges are yet to be placed.

Now for every $n_i, n_j \in N_1$, there is an edge $(n_i, n_j) \in E_1$ precisely when $n_i$ and $n_j$ are at consecutive depths (without loss of generality, say $n_i$ is at depth $\delta$ and $n_j$ is at depth $\delta + 1$) and $v_{i,j}(T_1) = \delta$, since this means that $n_i$ is the infector of $n_j$. Since $v_S(T_1) = v_S(T_2)$, we have that $v_{i,j}(T_1) = v_{i,j}(T_2)$ for all $n_i, n_j \in S$, and so $(n_i, n_j) \in E_1$ if and only if $(\phi(n_i), \phi(n_j)) = (n_i, n_j) \in E_2$. Thus $E_1 = E_2$ and $T_1$ and $T_2$ are $S$-isomorphic.



$$M = \begin{bmatrix} 0 & & & & & & \\ {\color{red}0} & 1 & & & & & \\ {\color{blue}0} & {\color{red}1} & 2 & & & & \\ {\color{blue}0} & {\color{red}1} & {\color{blue}1} & 2 & & & \\ {\color{blue}0} & {\color{blue}1} & {\color{red}2} & {\color{red}1} & 3 & & \\ {\color{blue}0} & {\color{blue}1} & {\color{red}1} & {\color{red}2} & {\color{blue}1} & 3 & \\ {\color{blue}0} & {\color{blue}1} & {\color{red}1} & {\color{red}2} & {\color{blue}1} & {\color{blue}2} & 3 \end{bmatrix}$$

FIG. 3. *For ease of visual notation, we have written the vector $v$ here as a matrix $M$, where $M_{i,j} = v_{i,j}$, and omitted the upper triangle of the matrix because $M$ is symmetric. The $v_{i,i}$ entries, shown in* (a) *as the black, diagonal entries of $M$, determine the depths of the nodes in the transmission tree. We place each node at its appropriate depth* (b). *Transmissions* (*directed edges*) *will be placed to point downwards, from one depth to the next. It then remains to check the* (*red*) *entries of $M$ corresponding to pairs of nodes at consecutive depths, in order to place the edges in the tree. To draw the transmission tree as a planar graph, it may be desirable to rearrange the order of the nodes at each depth*; *here, we have swapped the order of nodes* 3 *and* 4. *Blue entries of $M$ are not required for this tree reconstruction.*

Now suppose that $S$ is a strict subset of $N_1, N_2$ (there are some unsampled cases in each tree). By Lemma 1, we know that if there are any unsampled cases in $T_1$ and/or $T_2$ without sampled descendants, then these will not affect the vectors $v|_S(T_1), v|_S(T_2)$. It remains to show that there is a bijective function $\phi : N_1^* \to N_2^*$ such that $\phi$ is the identity on $S$ and $(n_i, n_j) \in E_1^*$ if and only if $\phi(n_i, n_j) \in E_2^*$.

If the source case in $T_1$ is an unsampled case, then $v_{s_i,s_i}(T_1) > 0$ for all $s_i \in S$. Since $v|_S(T_1) = v|_S(T_2)$, we also have $v_{s_i,s_i}(T_2) > 0$ for all $s_i \in S$, and so the source case is unsampled in $T_2$ also. From the first part of the proof, we know that any subtree (a connected subset of nodes) of sampled cases $\hat{S} \subseteq S$ which includes the source case must give rise to a unique vector $v|\hat{s}$, so that all node depths and edges are determined. By extension, any subtree $T|_{\hat{S}}$ of sampled cases $\hat{S}$ whose minimum depth in $T$ is $\delta$ must also be uniquely determined by $v|_{\hat{s}}$, since $v|_{\hat{s}}(T) = v|_{\hat{s}}(T|_{\hat{S}}) + \delta$. Therefore, we know that the identity map $\phi : S \to S$ preserves all edges within subtrees of sampled cases: for all $s_i, s_j \in S$, $(s_i, s_j) \in E_1^*$ if and only if $(s_i, s_j) \in E_2^*$.

It remains to show that for any path in $T_1$ from the source to a sampled case, an $S$-isomorphic path exists in $T_2$ (a path can be considered as a tree so we are continuing to use the same definition of $S$-isomorphism). By definition, the path $p_i$ from the source to a sampled node $n_i \in S$ at depth $\delta$ contains a single node at each depth $1, 2, \dots, \delta$, and recall that each sampled node has the same depth in $T_1$ and $T_2$.

Fix a sampled node $n_i$ at depth $\delta \geq 0$ and consider the path to it from the source case in each tree, $p_i(T_1)$ and $p_i(T_2)$ in $T_1$ and $T_2$, respectively. Consider a depth $x \in \{0, \dots, \delta\}$ and find the node $n_a$ at depth $x$ on $p_i(T_1)$. If $n_a$ is a sampled node, then $v_{a,a}(T_1) = x = v_{a,a}(T_2)$ and $v_{a,i}(T_1) = x = v_{a,i}(T_2)$, so the same sampled node $n_a$ also appears at depth $x$ on path $p_i(T_2)$. Now suppose that $n_a$ is an unsampled node in $T_1$, that is, $n_a \in U_1^*$. Since there is exactly one node at depth $x$ in $T_1$ which has $n_i$ among its descendants, and since this node $n_a$ is unsampled, then there can be no sampled node $n_b \in S$ such that both $v_{b,i}(T_1) = x$ and $v_{b,b}(T_1) = x$. Since the vectors are equal, there can be no node $n_c \in S$ such that both $v_{c,i}(T_2) = x$ and $v_{c,c}(T_2) = x$, and so the node at depth $x$ on path $p_i(T_2)$ is unsampled also.

Thus each edge $(n_a, n_b)$ on path $p_i(T_1)$ is in $E_1^*$ and has a corresponding edge $(\phi(n_a), \phi(n_b))$ on path $p_i(T_2)$ in $E_2^*$, where $n_a \in S$ if and only if $\phi(n_a) = n_a \in S$, and $n_a \in U_1^*$ if and only if $n_a \in U_2^*$; similarly for $n_b$. Since this is true for every path from the source to a sampled node, we have shown that $v|_S(T_1) = v|_S(T_2)$ implies that all such paths are $S$-isomorphic in $T_1$ and $T_2$, hence $T_1$ and $T_2$ are $S$-isomorphic. □

Note that this proof illustrates that many of the entries of $v$ are redundant for the reconstruction of the tree, particularly when all cases are sampled, in which case we can ignore any entries $v_{a,b}$ where $n_a$ and $n_b$ are not at consecutive depths. In Figure 3, we only need the diagonal and red entries of $M$ to construct the tree. In fact, since we know that the graph is a tree, wherever there are two red entries in a row, only one red entry is strictly necessary in this example for the placement of edges. Nevertheless, further (blue) entries are needed to understand the relationships across multiple depths when there are unsampled cases, and these "extra" entries also add weight in the comparison of transmission trees and may be useful if this metric was extended to include edge weights.

The existence of a metric on a set of objects enables a variety of further analyses to be performed. These include: visualising the pairwise distances between the objects using projections such as multi-dimensional scaling (MDS) [6] and cluster analysis, as proposed in the related literature of phylogenetic tree comparison [1, 16, 17, 5, 2, 22]. Although the metric we have proposed is not convex, barycentric methods can be used to find a representative "central" tree from a set, for example, we can find the geometric median tree as proposed in [22].

Such methods may be used to compare trees: from different input data, taking into account various combinations of metadata; from different inference processes, with variations in their assumptions and settings; and within the same inference process, for example, to assess convergence within a Bayesian posterior. Projecting tree–tree distances into two or three dimensions, assessing clustering and finding representative tree(s) can be important for assessing and summarising the performance of inference processes. Additionally, each tree can be compared to a fixed reference tree, for example, to assess the success of an inference process in reconstructing the "true" tree from a simulation, or to estimate the effective sample size of discrete tree structures as proposed in [25].

## 3. RESULTS

### 3.1 Toy Examples

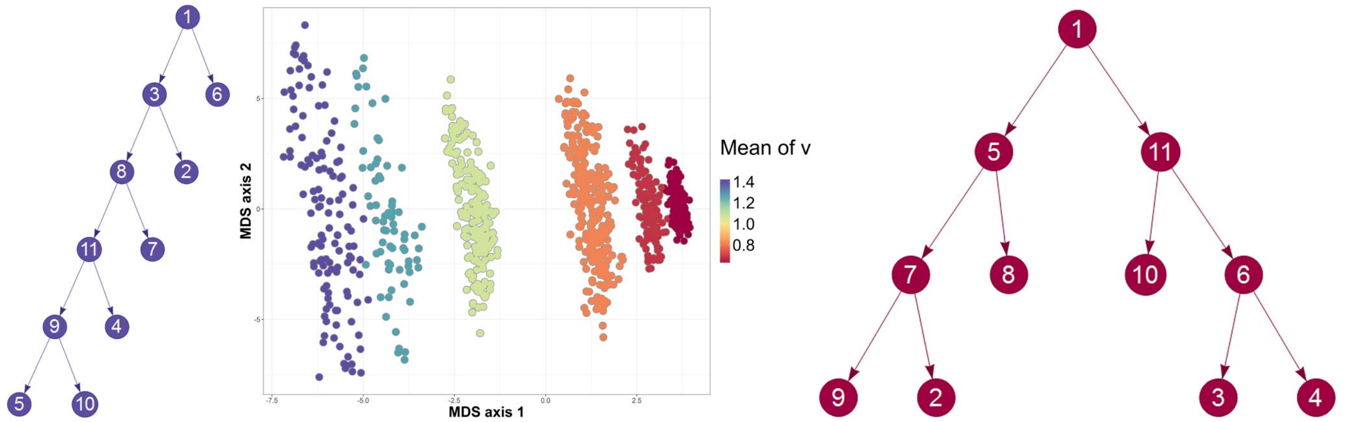
The metric which we have proposed here detects any differences between trees. In particular, it highlights differences in the "shape" of the transmission tree (star-like versus single transmission chain, etc.), corresponding to different transmission dynamics. The shape and depth of the tree is largely determined by the number of infectees per infector. The measure also highlights differences in the attribution of the source case (and in general, differences in historic transmissions are given more emphasis than recent transmission differences). The distance between two trees also depends on the number and relative positions of unsampled cases.

We tested how well the metric resolves some of these differences using small examples. For each of the following scenarios, we generated 1000 transmission trees at random from the set of trees with the given constraints. We then applied the metric to find the pairwise distances between them, and projected the distances into a two-dimensional plot using MDS. We use colours and shapes in the plots to highlight key differences between the trees, and to see where these colours do or do not correspond to position in the MDS. For each scenario, we picked the number of infected cases to be small enough so that it was easy to plot and examine the individual trees by eye, and for it to be possible to take a reasonably large sample from the set of all transmission trees of that size, but large enough for there to be a variety of possible tree structures within the given constraints.

3.1.1 *Scenario* 1. For the first scenario, we generated random transmission trees under the following constraints: we had exactly eleven sampled cases and no unsampled cases. Each infector was constrained to infect exactly two cases (a binary tree), and the source case was fixed as case 1. Under these constraints, there are precisely six possible tree "shapes", each admitting a variety of possible transmission trees through the permutation of the remaining ten case labels. The key variation in the MDS plot is associated with the height/shape of the tree: in Figure 4(b), we have coloured each point according to the mean value of its tree vector $v$, that is, the mean of the MRCI depths in the tree. Some example trees are also shown: Figure 4(a) is a tree with the maximum possible depth (mean of $v \approx 1.4$) and Figure 4(c) is a tree with the minimum possible depth (mean of $v \approx 0.6$), given the above constraints. The metric distinguishes trees composed of one long transmission chain in which each infection gives rise to only one onward-infecting case (and one case who does not infect anyone else), as opposed to more heterogeneous transmission trees in which some individuals cause two onward *infectious* cases.

3.1.2 *Scenario* 2. Our second scenario is similar to the first (eleven sampled cases, no unsampled cases, each infector infects exactly two cases), but now we fix the source case to be case 1 in half the trees, and case 2 in the other half. The resulting MDS plot is shown in Figure 5. The symmetry in the plot with respect to MDS axis 1 (which corresponds to the eigenvector with largest eigenvalue in the dimensionality reduction) illustrates symmetry in the tree distances with respect to the choice of source case (indicated by the shape of each point). The shape of the tree, as measured by the mean of $v$, varies strongly with the second MDS axis. Overall, Figure 5 illustrates that the metric is sensitive to both the shape and the labels in the transmission tree.

3.1.3 *Scenario* 3. We now reduce the constraint on the number of infectees. For our third scenario, each infector infects $n$ cases, where $n$ is picked uniformly at



(a) (b) (c)

FIG. 4. *Scenario* 1: *eleven sampled cases*, *no unsampled cases*, *each infector infects exactly two cases*, *source case fixed as case* 1.

random from $\{1, 2, 3\}$, per tree. Each tree has thirteen sampled cases and no unsampled cases. The source case is picked uniformly at random from $\{1, \ldots, 6\}$ (for ease of identification by colour in the MDS plots) and the remaining case labels are determined by a random permutation. The overwhelming grouping on the first two axes [Figure 6(a)] is by the number of infectees per infector. In particular, the transmission trees where each infector has one infectee, which are simple chains, are strongly separated from the other trees and are more widely spread in the plot. This is because the large number of possible permutations of their labels lead to greater differences in transmission histories than in the shorter, more balanced trees where each infector causes two or three new infections. There is still some noticeable separation by source case, which becomes much more apparent in a plot of the second and third axes [Figure 6(b)]. This underlines the findings of Scenarios 1 and 2 by showing that the metric distinguishes trees by transmission dynamics and source case attribution, but with rather more emphasis on the former when everything else is fixed.

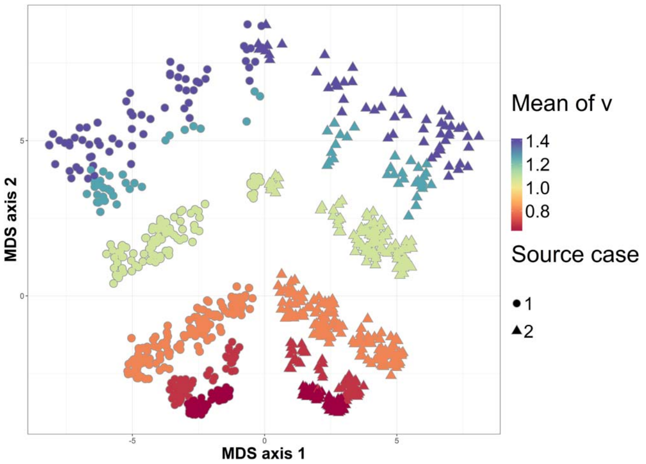
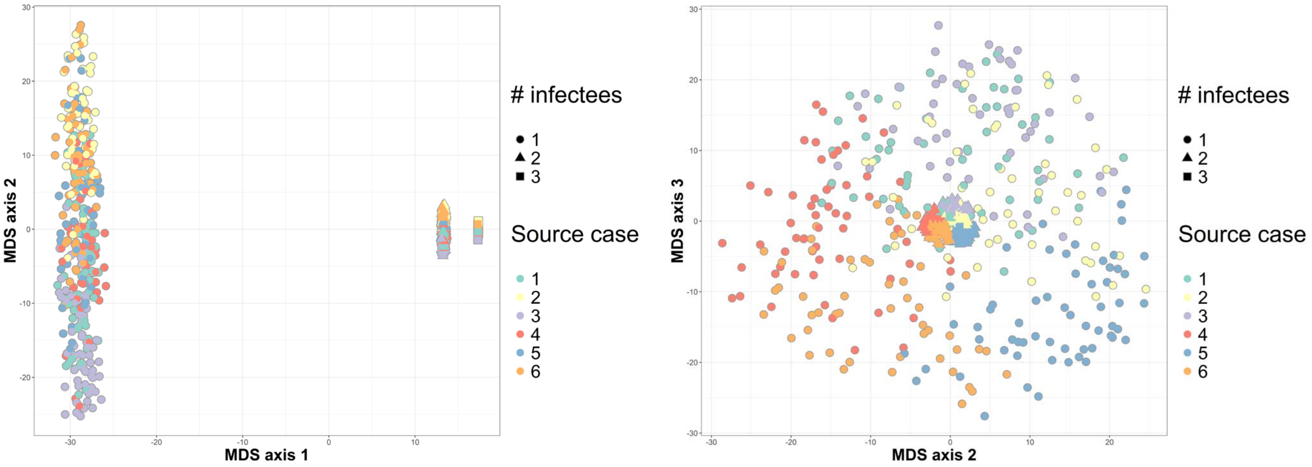


FIG. 5. *Scenario* 2: *similar to Scenario* 1, *except half have the source case fixed as case* 1, *the other half have the source case fixed as case* 2. *We note that the source cases are fixed in the trees* (*they are necessary to compute the distances*) *and are not revealed by the metric*.

3.1.4 *Scenario* 4. In our next scenario, we analyse the impact of including unsampled cases in our transmission tree. We consider trees with eight sampled cases and a further $c$ unsampled cases, where $c$ is picked uniformly at random from $\{0, \ldots, 8\}$. Each infector infects $n$ cases, where $n$ is picked uniformly at random from $\{2, \ldots, 6\}$, until all cases have been infected (note that this means that not every infector will necessarily infect *exactly* $n$ cases). Figure 7 shows how various characteristics of the transmission trees are represented in the MDS plot. The first two axes group the trees by features which are correlated with tree shape/transmission dynamics: the mean number of infectees per infector [Figure 7(a)] and the number of unsampled cases in the tree [Figure 7(b)]. These features are also strongly correlated with the mean of the tree vector $v|_S$ (which captures the depths of *sampled* MRCIs). As in Scenario 3, there is some grouping by source case [Figure 7(c)], particularly by *sampled* source case, especially in the second and third axes [Figure 7(d)], where we have plotted the trees with unsampled source cases with low point opacity.

3.1.5 *Scenario* 5. We compared trees with "super-spreaders" to those without. A "super-spreader" is an individual who infects a high number of secondary cases compared to other individuals. We simulate 300



(a) MDS plot of axes 1 and 2

(b) MDS plot of axes 2 and 3

FIG. 6. *Scenario* 3: *thirteen sampled cases*, *no unsampled cases*, *each infector infects n cases*, *where n is picked uniformly at random from* $\{1, 2, 3\}$, *per tree. Source case is picked uniformly at random from* $\{1, \ldots, 6\}$.

transmission trees with half of them containing a super-spreader. For each tree, there are 20 (sampled) cases from which a super-spreader was randomly chosen and can infect up to 10 cases, with the probability that it infects exactly 10 cases being 0.9; and the source case was fixed to be case 1. We find that the metric does not separate transmission trees with super-spreaders from those without, though super-spreader trees have a wider spread of tree–tree distances (and so visually occupy a larger region of the MDS space). Figure 8 illustrates the results. The lack of separation indicates that similar $v$ can be obtained from trees with widely varying maximum numbers of secondary infections.

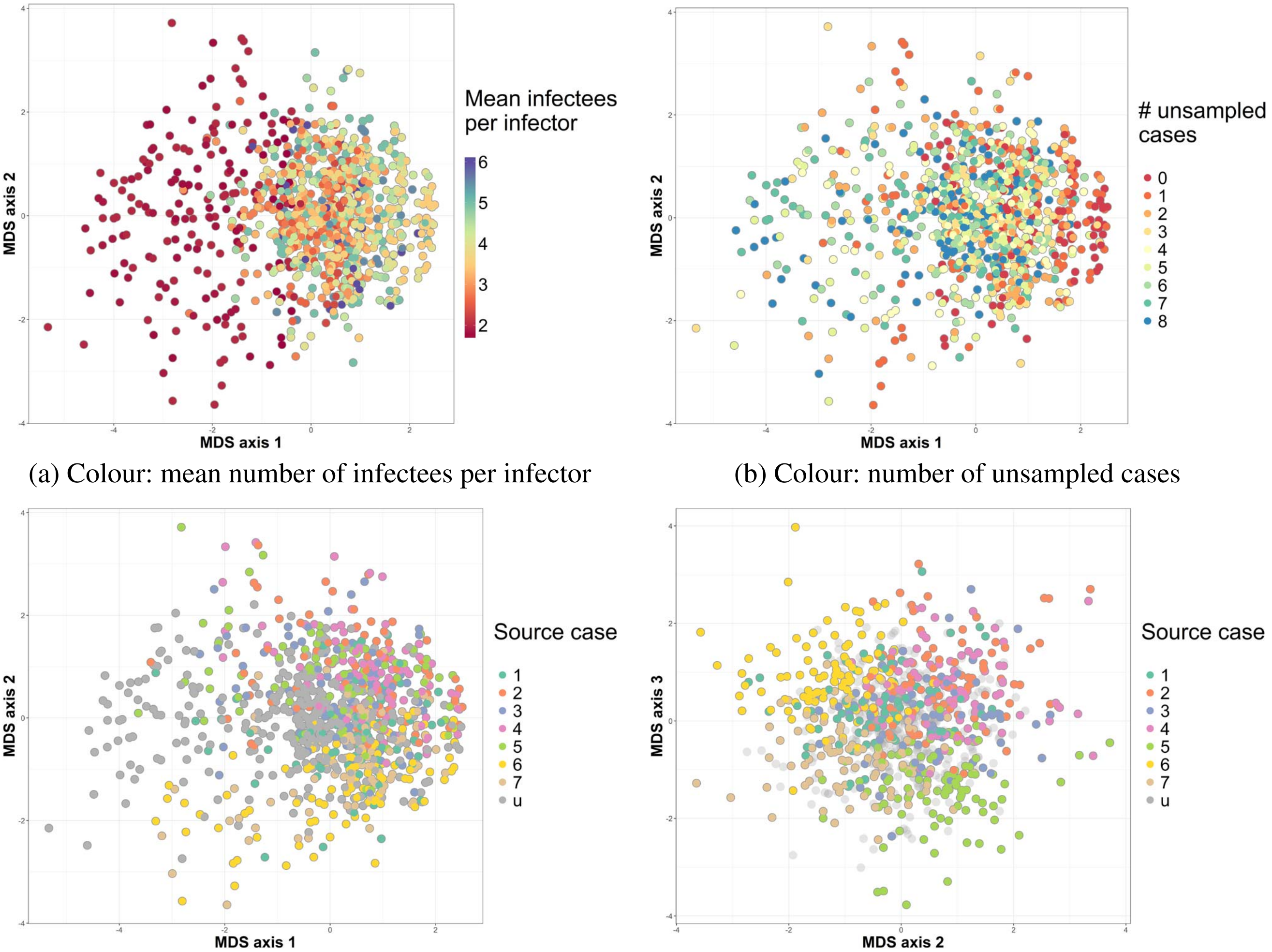
One observation that might explain the failure of the metric to distinguish a transmission tree containing a super-spreader (sp-tree) from one that does not (non-sp-tree) is that the tree vector of an sp-tree is closer to the line with slope one (in some space $\mathbb{R}^d$) than that of non-sp-tree. In fact, if the super-spreader in the sp-tree infects $n$ cases, there would be at least $\binom{n}{2}$ identical entries in the tree vector, being the depth of the common MRCI; and so the distance to the slope-one line would get smaller. Note that this does not necessarily imply that an sp-tree and a non-sp-tree are far apart from each other in the sense of the defined metric.

The wider tree–tree distance in the case of sp-trees can be explained by noting that infectees of a super-spreader occurring near the source (root) of the tree have much smaller depth of common MRCI than if the super-spreader were to occur far from the source.

### 3.2 Tuberculosis Outbreak

We used the R package TransPhylo [8] to perform MCMC inference to reconstruct an outbreak of tuberculosis (TB) reported by Roetzer et al. [32]. The outbreak lasted from 1997 to 2010 during which epidemological data were collected such as information concerning previous exposure to known cases, residence status, sex and age. TransPhylo is a Bayesian inference method to infer transmission trees using genomic data. TransPhylo's starting point is a timed phylogenetic tree, in which tips correspond to sampled cases and internal nodes correspond to inferred common ancestors; edge lengths are in units of time. The starting tree was inferred using the BEAST [10] software as described in [8]. This tree is held fixed, and TransPhylo proceeds by overlaying transmission events on it, and computing the likelihood of the overall transmission process at each iteration.

Here, we use the metric we have presented to compare inferred transmission trees under different priors, and to explore convergence of the MCMC. The time between an individual becoming infected and infecting others is a major source of uncertainty in TB, as it has a long and variable latent period; this is in contrast to acute infections such as influenza in which the generation time is short and not highly variable (typically under 1–2 weeks). In any public health investigation, it is difficult to determine how effectively and rapidly cases are identified. Accordingly, it is important to know how prior assumptions about these distributions affect outbreak reconstructions. The metric allows us to quantify and visualise this.



(a) Colour: mean number of infectees per infector

(b) Colour: number of unsampled cases

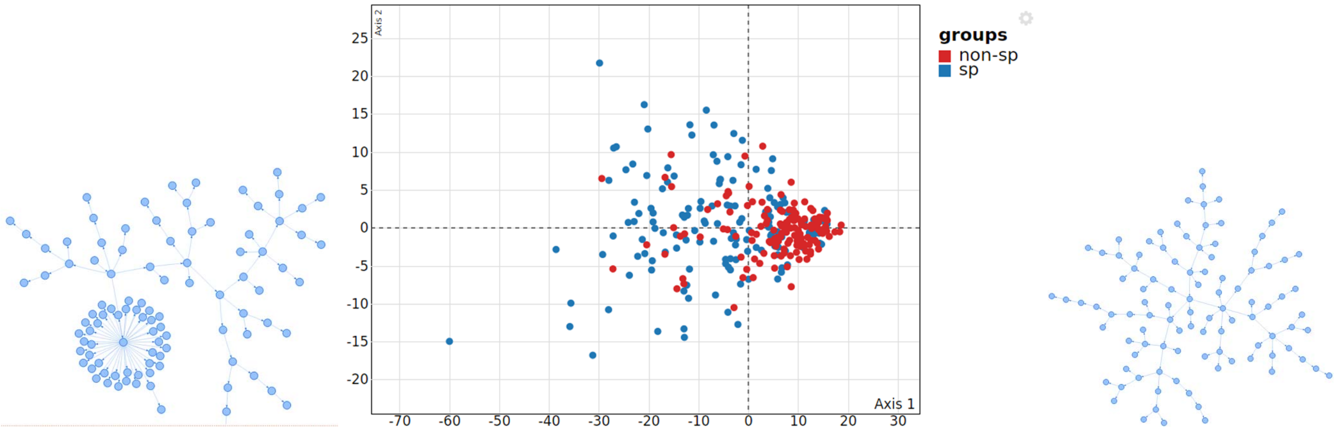
(c) Colour: ID of source case (or *u* for an unsampled source)

(d) Axes 2 and 3. Colour: ID of source case (or *u* for an unsampled source, faded)

FIG. 7. *MDS plots of tree–tree distances for trees from Scenario* 4: *eight sampled cases*, *up to eight further unsampled cases*, *each infector infects two to six cases*. *Colour is used to demonstrate how the trees are grouped according to various features*. *Axes* 1 *and* 2 *are plotted except where otherwise stated*.

We ran 100,000 MCMC iterations with five different choices for the priors for the sampling and generation times. Some individuals were sampled for reasons other than their symptoms and as such the prior sampling distribution was chosen to be a gamma distribution [8]. Also a gamma distribution was used for the prior generation time distribution in order to reflect the variable disease progression of TB. We sampled 200 random trees from the last 10,000 iterations of each of the five MCMC runs. We applied the metric to these trees and projected the distances into a two-dimensional plot using MDS (Figure 9). In Figure 9(a), we show the distances between the last 1000 trees from one of the MCMC runs, each tree colored by its iteration number. This reflects how the MCMC moves through the tree space: it samples several times from an area and then hops to another, qualitatively illustrating good mixing.

Figure 9 illustrates that there are distinct differences between the inferred trees depending on the priors. Figures 9(b) and 9(c) show 1000 trees, 200 from each of the five MCMC runs, on axes $1, 2$ and $2, 3$, respectively. Colors correspond to mean generation times and shape corresponds to mean sampling times. In Figure 9(b), there are two visually separated clusters of trees. It is not clear why the mean prior generation time of 4.3 years and sampling prior of 2.8 years should produce markedly different trees, as these are not extremal choices of the prior, but in practice it is useful to be able to visualise how unimodal a posterior (or set



(a) A super-spreader tree

(b) MDS plot

(c) A nonsuper-spreader tree

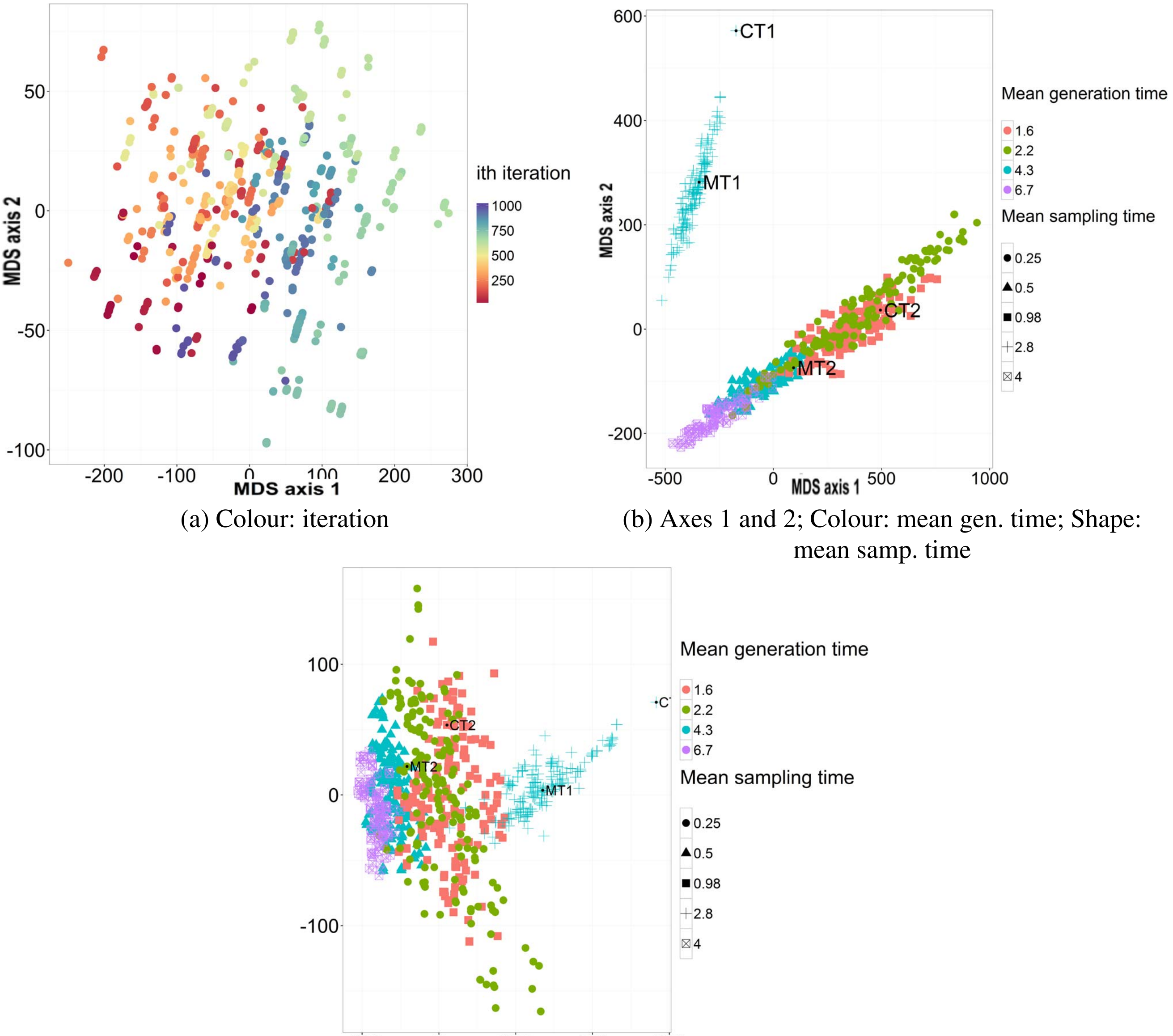
FIG. 8. *Transmission trees from a process with and without super-spreaders. In the MDS plot, trees with and without super-spreaders are colored in blue and red, respectively. The MDS plot suggests that the metric cannot separate the two groups, although the super-spreader group has a wider spread.*

of trees from multiple posteriors under different priors) is. For the two obvious clusters (blue, and everything else, in the middle panel of Figure 9), we obtain both a median tree using our metric and a consensus tree using TransPhylo's function consTTree which implements Edmond's algorithm. We refer to the smaller blue cluster as cluster 1 and the other as cluster 2. The points $MT1$, $MT2$ correspond to median trees for clusters 1 and 2 while $CT1$ and $CT2$ correspond to (Edmond's) consensus trees of these clusters. $CT1$ is visually separated from the rest of its cluster in the MDS plot, whereas the median trees sit centrally in their clusters. Consistent with this, the mean distances from MT1 and CT1 to trees in cluster 1 are 98 and 306 units, respectively. Cluster 2 is larger and more dispersed, and the consensus tree is more central, but the mean distances between MT2 and CT2 and cluster 2's trees are 370 versus 474 units. In our metric, the median trees are closer to the clusters they aim to summarise than the trees derived from Edmond's algorithm. The individual transmission trees are illustrated in Figure 10.

Trees from the two main clusters have similar depths, and all identify case 1 as the source. Trees from within each cluster have strong similarity in the first few infections after the source case, but there are distinct differences between the clusters, with many individuals placed very differently. For example, note the positions of patients 83 and 85, who appear early in the transmission process in cluster 1 but at the end, with no infectees, in cluster 2. Overall, trees from cluster 2 have more unsampled cases (average 88) than cluster 1 (average 33). This is reflected in the median and consensus trees, with 38 and 60 unsampled cases in MT1 and CT1 versus 145 and 111 in MT2 and CT2, respectively. This is likely a result of the prior assumptions: shorter sampling and generation times (more in cluster 2) use higher numbers of unsampled cases to fill in transmission events along long branches of the fixed phylogenetic tree that is provided as input.

We visualised the median and consensus trees using colour to indicate patients' TB smear status. The smear status refers to the result of a sputum smear microscopy test, which detects TB bacilli in patient sputum samples. Smear-positive individuals are believed to transmit TB more than smear-negative cases due to the higher numbers of bacilli present in the sputum [34], but the smear test itself has limited sensitivity (as low as 50%) [33]. In our analysis, smear-positive individuals transmit more in trees MT1 and CT1 than in MT2 and CT2, largely due to the fact that MT2 and CT2 have a much higher fraction of transmission by unsampled cases.

The metric can be used to compare analyses of the same dataset with different inference methods, which have different underlying assumptions, constraints and priors. We compared four methods, analysing the tuberculosis outbreak data with each (Figure 11). Beastlier and phybreak seem closest in the visualisation; both simultaneously estimate the phylogenetic and transmission trees and do not allow unsampled cases. SCOTTI's approach requires the exposure times for all



(a) Colour: iteration

(b) Axes 1 and 2; Colour: mean gen. time; Shape: mean samp. time

(c) Axes 2 and 3; Colour: mean gen. time; Shape: mean samp. time

FIG. 9. *MDS plots of tree–tree distances for posterior transmission trees from the Hamburg TB outbreak* [32]. (a) *Colour indicates iteration number in the MCMC chain*. (b) *Colour indicates mean prior generation time*, *shape indicates mean prior sampling time and the median trees of the two groups are labelled MT*1 *and MT*2.

the cases (we do not know these so for the purposes of demonstration, we simulated them), and the unsampled state in SCOTTI is more appropriate for an environmental pathogen than for an unsampled human host. SCOTTI is based on the structured coalescent, with constant rates of migration of lineages among demes (here, hosts). SCOTTI is therefore quite unlike the other approaches. Finally, TransPhylo uses a single input timed phylogenetic tree, and allows for unsampled cases, which likely accounts for its distance to the trees estimated by phybreak and Beastlier.

The data behind Figure 11 are based on model configurations which were kept as consistent as their differences allow. For example, generation time priors are Gamma distributed with identical parameters for the TransPhylo, Beastlier and phybreak models, whereas for SCOTTI these are pre-generated (we used a statistical model of the time between infection and sampling and the known sampling dates; this was the same

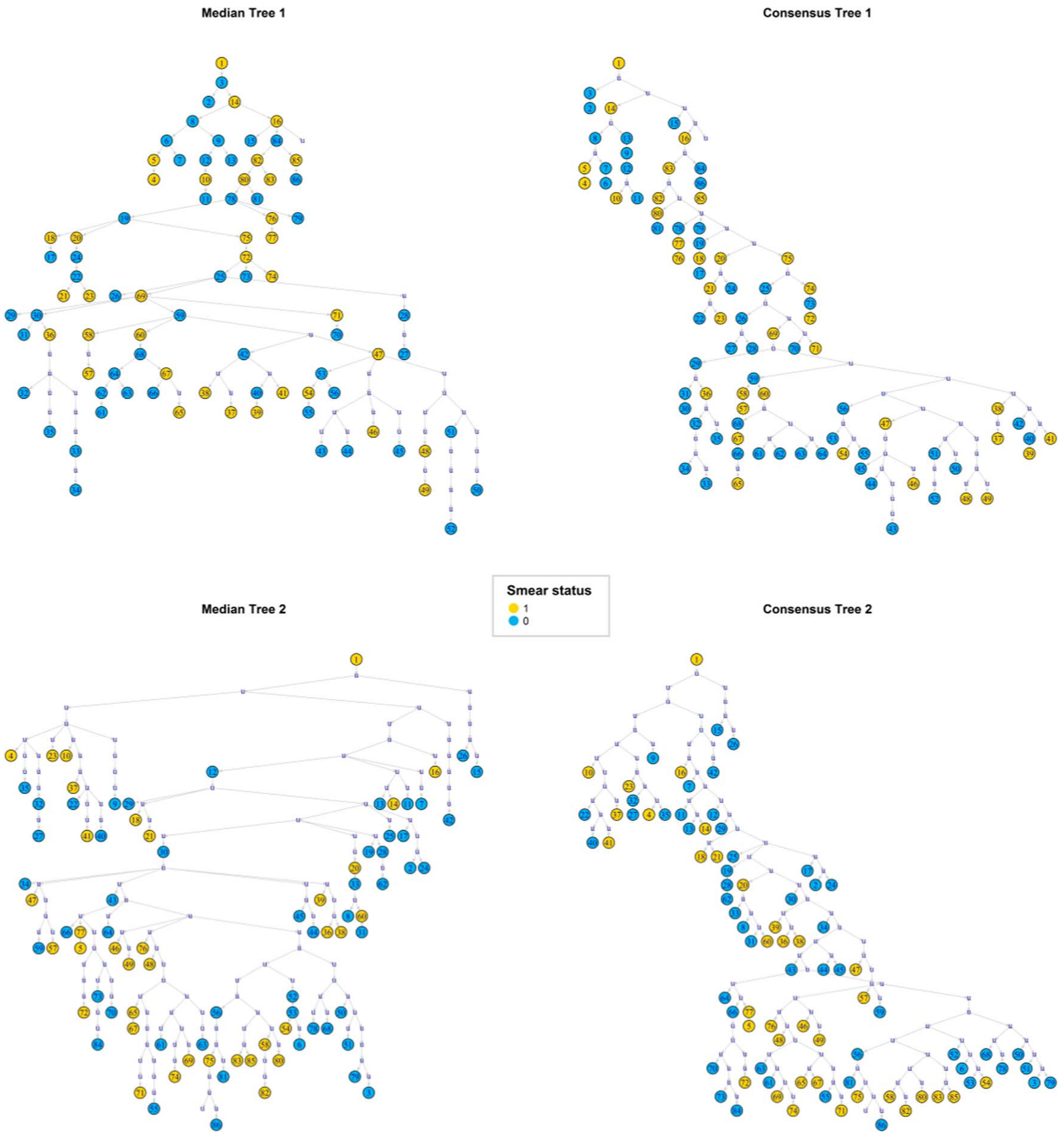


FIG. 10. *Median and consensus trees from each of the two clusters, coloured according to the smear status of each sampled patient.*

gamma distributions used in TransPhylo) and passed in as fixed periods. Similarly, sample time priors can only be specified for TransPhylo and phybreak. 100,000 simulations were run for the SCOTTI and TransPhylo data, with 20,000 for Beastlier and phybreak.

## 4. DISCUSSION

We have introduced a metric, in the sense of a true distance function, on the set of transmission trees with labelled sampled cases along with unsampled cases (up to our notion of isomorphism). In the context of inferring transmission trees, this metric can aid in assessing convergence, posterior concordance and sensitivity to priors, and in comparing inference methods to each other. It emphasises the source case and the extent of shared transmission events in two trees. We applied the metric to random trees from simple simulated scenarios and found that it can separate trees according to their overall shape, the numbers of infectees per infector, and according to which case is the source. It allows for trees with unsampled cases, an advantage because health authorities rarely know about every case in an outbreak of an infectious disease.

The metric is sensitive to the source case, and as such, it carries the limitation that trees with different

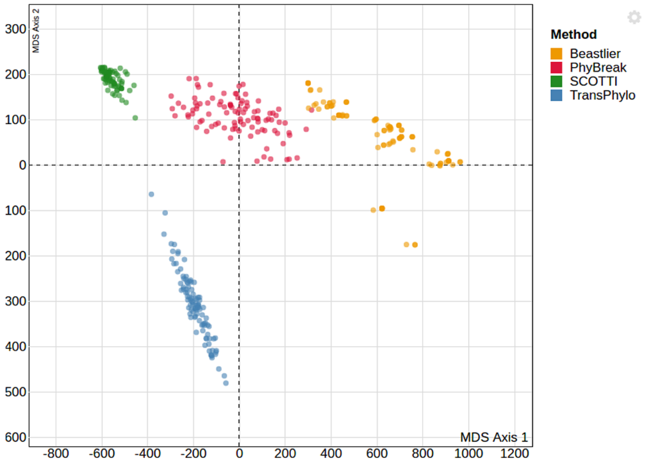


FIG. 11. *MDS plot of transmission trees estimated by TransPhylo, SCOTTI, phybreak and Beastlier.*

source cases but otherwise similar transmission events may appear a higher distance from each other than intuition would suggest. In addition, while unsampled cases are possible, the metric is only a metric up to pruning of unsampled cases with no descendants, and up to relabelling of unsampled cases. The way we treat unsampled cases could result in distances that do not always reflect intuition. For example, if one tree has long chains of unsampled cases but otherwise similar connectivity (i.e., A infects B, versus A infects B via a long chain of intermediate unsampled cases, and this occurs for many pairs of individuals), our metric will show a relatively large distance. If this is not desired in a specific application, the effect can be reduced by collapsing chains of unsampled cases before computing distances.

The metric as it stands also does not take the timing of transmission events into account, equating for example a tree in which A infects B and then infects C two weeks later, with one in which A infects C and then infects B a year later (as both have A infecting both B and C). It would be straightforward, however, to modify the metric in either of two ways: (1) convert the transmission tree to a genealogical, binary, tree—capturing pathogen lineages that branch at transmission events—and then use a metric on those binary trees [31, 3, 22], or (2) incorporate timing information in the lengths of branches in the framework we have presented here. In (2), we would construct a vector $w_S(T)$ whose entries were the *time elapsed* between the infection of the MRCIs, rather than the *depths* of the MRCIs, and then the time-sensitive metric could be defined as

$$\begin{aligned} d(T_1, T_2) = \big\| &\big(\varepsilon v|_S(T_1) + (1-\varepsilon) w|_S(T_1)\big) \\ &- \big(\varepsilon v|_S(T_2) + (1-\varepsilon) w|_S(T_2)\big)\big\|. \end{aligned}$$

With $\varepsilon > 0$, this would still be a metric on $\mathcal{T}$ up to the same isomorphism.

The metric could be used in other applications analogous to those for phylogenetic trees. For example, Nye et al. created parsimonious meta-trees to capture the relationships among a set of phylogenetic trees, scoring each meta-tree with the Robinson–Foulds metric [29]. The same approach could be taken here to create a meta-tree of transmission trees. The metric could also be used to aid in computing effective sample sizes for posterior collections of transmission trees. Effective sample sizes (ESS) are routinely used in phylogenetic inference, and should be adopted for inference of transmission trees as well. Recently, Lanfear et al. [25] outlined approaches to use distances been phylogenetic tree topologies to compare MCMC runs and assess convergence and autocorrelation—they used traces of distances between trees along the MCMC chains and a single "focal tree", and distances between trees in the chain sampled at different sampling intervals ("jump distances"). Lanfear et al. computed effective sample sizes by applying standard techniques to distances between posterior trees. The same approaches could be used to estimate effective sample sizes for MCMC chains inferring transmission trees, using the metric we have presented here.

## 5. CONCLUDING REMARKS

Inferring transmission events from epidemiological, clinical and now genetic data is a challenging task, and an important one as understanding transmission is essential for designing the best approaches to control infections. Genomic data are noisy, and the underlying processes generating the true variation are stochastic. However, recent advances in sequencing technologies have led to widespread interest in using pathogen sequences to inform us about who infected whom. There are now many Bayesian methods available for this inference task, each developed with specific goals and features in mind, and each tested on the authors' own data and simulation scenario (with [23] as one exception that includes tests on other authors' simulations).

Understanding convergence, the effects of priors and the structure of the posterior collections of transmission trees is not trivial. As this field matures, comparing and benchmarking the performance of different methods will require the ability to quantify how close different approaches come to each other and to gold standard trees that experts agree are the best match to comprehensive data sources for an outbreak. We have developed a metric that can aid in these tasks, illustrated its performance and made it available to the community.

## 6. AVAILABILITY

The R functions required for the tree distances presented here are available in the `treespace` package [19, 20]. A worked example for transmission trees is available on the `treespace` CRAN page: https://cran.r-project.org/web/packages/treespace/vignettes/TransmissionTreesVignette.html.

## ACKNOWLEDGEMENTS

This work was supported by the Engineering and Physical Sciences Research Council grants EP/K026003/1 (CC), EP/N014529/1 (CC and YX). JS is hosted by the Centre for Integrative Systems Biology and Bioinformatics at Imperial College London, UK. DA is funded by the Nigerian Universities Commission [under the presidential special scholarship scheme for innovation and development (PRESSID)].

## REFERENCES

[1] Amenta, N. and Klingner, J. (2002). Case study: Visualizing sets of evolutionary trees. In *IEEE Symposium on Information Visualization*, 2002. (*InfoVis'*02) 71–74.

[2] Berglund, D. (2011). Visualization of phylogenetic tree space. Ph.D. thesis, Stockholm Univ.

[3] Billera, L. J., Holmes, S. P. and Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.* **27** 733–767. MR1867931

[4] Cardona, G., Mir, A., Rossello Llompart, F., Rotger, L. and Sanchez, D. (2013). Cophenetic metrics for phylogenetic trees, after Sokal and Rohlf. *BMC Bioinform.* **14** 3.

[5] Chakerian, J. and Holmes, S. (2012). Computational tools for evaluating phylogenetic and hierarchical clustering trees. *J. Comput. Graph. Statist.* **21** 581–599. MR2970909

[6] Cox, T. F. and Cox, M. A. A. (2000). *Multidimensional Scaling*. CRC Press Boca Raton, FL.

[7] De Maio, N., Wu, C.-H. and Wilson, D. J. (2016). SCOTTI: Efficient reconstruction of transmission within outbreaks with the structured coalescent. *PLoS Comput. Biol.* **12** e1005130.

[8] Didelot, X., Fraser, C., Gardy, J. and Colijn, C. (2017). Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Mol. Biol. Evol.* **34** 997–1007.

[9] Didelot, X., Gardy, J. and Colijn, C. (2014). Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol. Biol. Evol.* **31** 1869–1879.

[10] Drummond, A. J. and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7** 214.

[11] Gardy, J., Loman, N. J. and Rambaut, A. (2015). Real-time digital pathogen surveillance—The time is now. *Genome Biol.* **16** 155.

[12] Gardy, J. L., Johnston, J. C., Ho Sui, S. J., Cook, V. J., Shah, L., Brodkin, E., Rempel, S., Moore, R., Zhao, Y., Holt, R., Varhol, R., Birol, I., Lem, M., Sharma, M. K., Elwood, K., Jones, S. J. M., Brinkman, F. S. L., Brunham, R. C. and Tang, P. (2011). Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N. Engl. J. Med.* **364** 730–739.

[13] Gibbons, A. (1985). *Algorithmic Graph Theory*. Cambridge Univ. Press, Cambridge.

[14] Gray, R. R., Tatem, A. J., Johnson, J. A., Alekseyenko, A. V., Pybus, O. G., Suchard, M. A. and Salemi, M. (2011). Testing spatiotemporal hypothesis of bacterial evolution using methicillin-resistant Staphylococcus aureus ST239 genome-wide data within a Bayesian framework. *Mol. Biol. Evol.* **28** 1593–1603.

[15] Hall, M., Woolhouse, M. and Rambaut, A. (2015). Epidemic reconstruction in a phylogenetics framework: Transmission trees as partitions of the node set. *PLoS Comput. Biol.* **11** e1004613.

[16] Hillis, D. M., Heath, T. A. and St John, K. (2005). Analysis and visualization of tree space. *Syst. Biol.* **54** 471–482.

[17] Holmes, S. (2006). Visualising data. In *Statistical Problems in Particle Physics, Astrophysics and Cosmology, Proceedings of PHYSTAT*05 (L. Lyons and M. K. Ünel, eds.) 197–208. Imperial College Press, London.

[18] Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C. and Ferguson, N. (2014). Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput. Biol.* **10** e1003457.

[19] Jombart, T., Kendall, M., Almagro-Garcia, J. and Colijn, C. (2017). treespace: Statistical exploration of landscapes of phylogenetic trees. R package version 1.0.0.

[20] Jombart, T., Kendall, M., Almagro-Garcia, J. and Colijn, C. (2017). treespace: Statistical exploration of landscapes of phylogenetic trees. *Mol. Ecol. Resour.* **17** 1385–1392.

[21] Kenah, E., Britton, T., Halloran, M. E. and Longini, I. M. Jr. (2016). Molecular infectious disease epidemiology: Survival analysis and algorithms linking phylogenies to transmission trees. *PLoS Comput. Biol.* **12** e1004869.

[22] Kendall, M. and Colijn, C. (2016). Mapping phylogenetic trees to reveal distinct patterns of evolution. *Mol. Biol. Evol.* **33** 2735–2743.

[23] Klinkenberg, D., Backer, J. A., Didelot, X., Colijn, C., Wallinga, J. and Haydon, D. T. (2017). Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput. Biol.* **13** e1005495.

[24] Köser, C. U., Holden, M. T. G., Ellington, M. J., Cartwright, E. J. P., Brown, N. M., Ogilvy-Stuart, A. L., Hsu, L. Y., Chewapreecha, C., Croucher, N. J., Harris, S. R., Sanders, M., Enright, M. C., Dougan, G., Bentley, S. D., Parkhill, J., Fraser, L. J., Betley, J. R., Schulz-Trieglaff, O. B., Smith, G. P. and Peacock, S. J. (2012). Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N. Engl. J. Med.* **366** 2267–2275.

[25] LANFEAR, R., HUA, X. and WARREN, D. L. (2016). Estimating the effective sample size of tree topologies from Bayesian phylogenetic analyses. *Genome Biol. Evol.* **8** 2319–2332.

[26] MOLLENTZE, N., NEL, L. H., TOWNSEND, S., LE ROUX, K., HAMPSON, K., HAYDON, D. T. and SOUBEYRAND, S. (2014). A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proc. R. Soc. Lond., B Biol. Sci.* **281** 20133251.

[27] MORELLI, M. J., THÉBAUD, G., CHADŒUF, J., KING, D. P., HAYDON, D. T. and SOUBEYRAND, S. (2012). A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Comput. Biol.* **8** e1002768.

[28] NUMMINEN, E., CHEWAPREECHA, C., SIRÉN, J., TURNER, C., TURNER, P., BENTLEY, S. D. and CORANDER, J. (2014). Two-phase importance sampling for inference about transmission trees. *Proc. R. Soc. Lond., B Biol. Sci.* **281** 20141324.

[29] NYE, T. M. W. (2008). Trees of trees: An approach to comparing multiple alternative phylogenies. *Syst. Biol.* **57** 785–794.

[30] QUICK, J., LOMAN, N. J., DURAFFOUR, S., SIMPSON, J. T., SEVERI, E., COWLEY, L., BORE, J. A., KOUNDOUNO, R., DUDAS, G., MIKHAIL, A., OUÉDRAOGO, N., AFROUGH, B., BAH, A., BAUM, J. H. J., BECKER-ZIAJA, B., BOETTCHER, J. P., CABEZA-CABRERIZO, M., CAMINO-SÁNCHEZ, Á., CARTER, L. L., DOERRBECKER, J., ENKIRCH, T., GARCÍA-DORIVAL, I., HETZELT, N., HINZMANN, J., HOLM, T., KAFETZOPOULOU, L. E., KOROPOGUI, M., KOSGEY, A., KUISMA, E., LOGUE, C. H., MAZZARELLI, A., MEISEL, S., MERTENS, M., MICHEL, J., NGABO, D., NITZSCHE, K., PALLASCH, E., PATRONO, L. V., PORTMANN, J., REPITS, J. G., RICKETT, N. Y., SACHSE, A., SINGETHAN, K., VITORIANO, I., YEMANABERHAN, R. L., ZEKENG, E. G., RACINE, T., BELLO, A., SALL, A. A., FAYE, O., FAYE, O., MAGASSOUBA, N., WILLIAMS, C. V., AMBURGEY, V., WINONA, L., DAVIS, E., GERLACH, J., WASHINGTON, F., MONTEIL, V., JOURDAIN, M., BERERD, M., CAMARA, A., SOMLARE, H., CAMARA, A., GERARD, M., BADO, G., BAILLET, B., DELAUNE, D., NEBIE, K. Y., DIARRA, A., SAVANE, Y., PALLAWO, R. B., GUTIERREZ, G. J., MILHANO, N., ROGER, I., WILLIAMS, C. J., YATTARA, F., LEWANDOWSKI, K., TAYLOR, J., RACHWAL, P., TURNER, D. J., POLLAKIS, G., HISCOX, J. A., MATTHEWS, D. A., O'SHEA, M. K., JOHNSTON, A. M., WILSON, D., HUTLEY, E., SMIT, E., DI CARO, A., WÖLFEL, R., STOECKER, K., FLEISCHMANN, E., GABRIEL, M., WELLER, S. A., KOIVOGUI, L., DIALLO, B., KEÏTA, S., RAMBAUT, A., FORMENTY, P., GÜNTHER, S. and CARROLL, M. W. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530** 228–232.

[31] ROBINSON, D. F. and FOULDS, L. R. (1979). Comparison of weighted labelled trees. In *Combinatorial Mathematics, VI* (*Proc. Sixth Austral. Conf., Univ. New England, Armidale*, 1978). *Lecture Notes in Math.* **748** 119–126. Springer, Berlin. MR0558039

[32] ROETZER, A., DIEL, R., KOHL, T. A., RÜCKERT, C., NÜBEL, U., BLOM, J., WIRTH, T., JAENICKE, S., SCHUBACK, S., RÜSCH-GERDES, S., SUPPLY, P., KALINOWSKI, J. and NIEMANN, S. (2013). Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: A longitudinal molecular epidemiological study. *PLoS Med.* **10** e1001387.

[33] SIDDIQI, K., LAMBERT, M.-L. and WALLEY, J. (2003). Clinical diagnosis of smear-negative pulmonary tuberculosis in low-income countries: The current evidence. *Lancet, Infect. Dis.* **3** 288–296.

[34] SINGH, M., MYNAK, M. L., KUMAR, L., MATHEW, J. L. and JINDAL, S. K. (2005). Prevalence and risk factors for transmission of infection among children in household contact with adults having pulmonary tuberculosis. *Arch. Dis. Child.* **90** 624–628.

[35] SOUBEYRAND, S. (2016). Construction of semi-Markov genetic-space-time SEIR models and inference. *J. Soc. Fr. Stat.* **157** 129–152.

[36] STADLER, T. and BONHOEFFER, S. (2013). Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **368** 20120198.

[37] WALKER, T. M., IP, C. L. C., HARRELL, R. H., EVANS, J. T., KAPATAI, G., DEDICOAT, M. J., EYRE, D. W., WILSON, D. J., HAWKEY, P. M., CROOK, D. W., PARKHILL, J., HARRIS, D., WALKER, A. S., BOWDEN, R., MONK, P., SMITH, E. G. and PETO, T. E. A. (2013). Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: A retrospective observational study. *Lancet, Infect. Dis.* **13** 137–146.

[38] WORBY, C. J., O'NEILL, P. D., KYPRAIOS, T., ROBOTHAM, J. V., DE ANGELIS, D., CARTWRIGHT, E. J. P., PEACOCK, S. J. and COOPER, B. S. (2016). Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *Ann. Appl. Stat.* **10** 395–417. MR3480501