Transmission trees on a known pathogen phylogeny: enumeration and sampling

Matthew Hall,*,[1] and Caroline Colijn[2]
[1]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, United Kingdom
[2]Department of Mathematics, Simon Fraser University, Canada
***Corresponding author:** E-mail: matthew.hall@bdi.ox.ac.uk.
Associate Editor:

Abstract

One approach to the reconstruction of infectious disease transmission trees from pathogen genomic data has been to use a phylogenetic tree, reconstructed from pathogen sequences, and annotate its internal nodes to provide a reconstruction of which host each lineage was in at each point in time. If only one pathogen lineage can be transmitted to a new host (i.e. the transmission bottleneck is complete), this corresponds to partitioning the nodes of the phylogeny into connected regions, each of which represents evolution in an individual host. These partitions define the possible transmission trees that are consistent with a given phylogenetic tree. However, the mathematical properties of the transmission trees given a phylogeny remain largely unexplored. Here, we describe a procedure to calculate the number of possible transmission trees for a given phylogeny, and we show how to uniformly sample from these transmission trees. The procedure is outlined for situations where one sample is available from each host and trees do not have branch lengths, and we also provide extensions for incomplete sampling, multiple sampling, and the application to time trees in a situation where limits on the period during which each host could have been infected are known. The sampling algorithm is available as an `R` package (STraTUS).

Key words: epidemic reconstruction, molecular epidemiology, pathogen genomics, phylogenetics

Article

Introduction

The use of genetic data to reconstruct a pathogen transmission tree (a graph representing who infected who in an epidemic) has been the subject of considerable interest in recent years. Many different approaches have been proposed, both phylogenetic and non-phylogenetic (Aldrin *et al.*, 2011; Jombart *et al.*, 2014; Skums *et al.*, 2017). In phylogenetic approaches, a phylogenetic tree reconstructed from sequences for pathogens sampled in an epidemic will specify the order of the coalescences of lineages, and also, if its nodes are dated, the time at which these occurred. Some approaches further assume that internal nodes in the phylogeny correspond to transmission events (Lau *et al.*, 2015; Mollentze *et al.*, 2014; Morelli *et al.*, 2012), which in a dated phylogeny specifies infection dates, while others do not (Didelot *et al.*, 2014, 2017; Hall *et al.*, 2015; Klinkenberg *et al.*, 2017). In either case, a phylogeny on its own does not determine who infected who,

and extra components are required to reconstruct transmission events.

The assumption of coinciding lineage coalescences and transmission events may be unwise, and in particular it does not take into account within-host pathogen diversity (Giardina *et al.*, 2017; Ypma *et al.*, 2013). Several approaches have been taken that do not make it, one of which is to note that if a phylogeny from a completely sampled outbreak has its nodes annotated with the hosts in which each lineage was present, the transmission tree is known (Didelot *et al.*, 2014, 2017; Hall *et al.*, 2015). In particular, Hall *et al.* (2015) noted that the set of transmission trees for a known phylogeny, with complete sampling and assuming transmission is a complete bottleneck, is equivalent to the set of partitions of its nodes with the property that each part of each partition contains at least one tip and the subgraph induced by the nodes in each part is connected. However, the mathematical properties of this space of partitions remain largely unexplored.

Here, we provide procedures for counting the total number of these partitions (and hence the total number of transmission trees) for a known phylogeny. We also give an algorithm that samples uniformly from the set of such partitions. Initially we assume that the phylogeny is binary, sampling is complete, that each host provided one sample, and that nothing is known about the timings of each infection, but we go on in the Appendix to individually relax each of these assumptions individually, and then relax them together.

The procedures outlined here may be useful to researchers wishing to explore the structure that the phylogeny imposes on transmission tree space, or alternatively to explore whether a candidate transmission event is firmly (mathematically) ruled out by a phylogeny or set of phylogenies. Uniform sampling from transmission trees on a phylogeny is rapid and could allow public health researchers who are reconstructing outbreaks a quick guide to some of the most frequently-occurring transmission events among all transmission trees consistent with a set of sequence data. We include some numerical applications of our sampling approach, comparing transmission trees on balanced vs unbalanced phylogenies, and comparing random uniformly sampled transmission trees to transmission trees inferred with the *TransPhylo* approach (Didelot *et al.*, 2017).

New Approaches

Here we describe how to count, and uniformly sample, transmission trees for a known phylogeny in the simplest case where the phylogeny is binary, each host in the transmission tree is sampled once and only once, and no time limits are placed on the duration of an infection.

Let the phylogeny $\mathcal{T}$ be an unlabelled rooted binary tree, without branch lengths. Let $\mathcal{T}^*$ represent the unrooted tree obtained from $\mathcal{T}$ by attaching a single extra tip to the root of $\mathcal{T}$ by

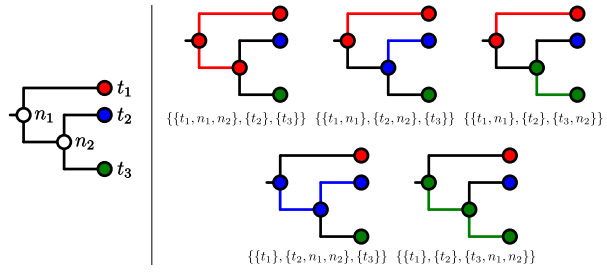

FIG. 1. A rooted phylogeny (left) and the five compatible transmission trees labelled with their expression as partitions of its node set (right). Coloured branches connect members of the same part.

a single edge. Note that two distinct $\mathcal{T}$s can have the same $\mathcal{T}^*$, and that $\mathcal{T}^*$ has one more tip than $\mathcal{T}$.

We follow the correspondence described by Hall *et al.* (2015) between transmission trees and partitions of the node set of $\mathcal{T}$ such that all tips derived from the same host are members of the same part (or block, or subset) of the partition, and the subgraph induced by each part is connected. This assumes that sampling is complete and that transmission is a complete bottleneck (i.e. that only one pathogen is transmitted at a time, so that diversity is not transmitted from host to host). While we relax the former assumption in the Appendix, the latter is more fundamental. See figure 1 for an example. We call a partition that satisfies these constraints an *admissible partition*.

In this paper, the term "subtree" is intended in the normal phylogenetic (rather than graph theoretic) sense: a subtree is a subgraph of $\mathcal{T}$ consisting of a node u, all its descendants (if any), and the edges between them. We denote the subtree rooted at u by $\mathcal{T}_u$; this is defined even if u is a tip.

Enumeration of possible transmission trees

With $\mathcal{T}$ fixed and having n tips, suppose we wish to count the number of admissible partitions, as defined above, of its node set $N(\mathcal{T})$, and hence the set of possible transmission trees. If the set of such partitions is $\mathbf{P}(\mathcal{T})$, we wish to calculate $|\mathbf{P}(\mathcal{T})|$. Nothing about the definition of an admissible partition requires a rooted tree, so $\mathbf{P}(\mathcal{T}^*)$ is defined similarly. It is trivial that if $n=1$ then $|\mathbf{P}(\mathcal{T})|=|\mathbf{P}(\mathcal{T}^*)|=1$. From here on, when we discuss partitions we mean admissible partitions.

If $\mathcal{T}_u$ is a subtree, we can define $\mathbf{P}(\mathcal{T}_u)$ in the obvious way by regarding $\mathcal{T}_u$ as a tree in its own right. If $\mathcal{T}_u$ is indeed a subtree in a larger phylogeny of an epidemic, however, this is not sufficient. We do not assume that transmission occurs at the time of internal nodes, and so, even with complete sampling, it is possible that the root node of any subtree was not infecting any of the hosts from which the tips of that subtree were sampled.

To allow for this possibility, we also define a second set of partitions of $N(\mathcal{T}_u)$, $\mathbf{Q}(\mathcal{T}_u)$:

$$\mathbf{Q}(\mathcal{T}_u)=\{\{N(\mathcal{T}_u)\cap S : S\in\mathfrak{P}\} : \mathfrak{P}\in\mathbf{P}(\mathcal{T})\}$$

An element of $\mathbf{Q}(\mathcal{T}_u)$ is the image of an element of $\mathbf{P}(\mathcal{T})$ when the intersections of all its parts with the node set of $\mathcal{T}_u$ are taken. (This is not an injective operation, as the partition of the nodes of $\mathcal{T}$ that are not nodes of $\mathcal{T}_u$ does not matter.)

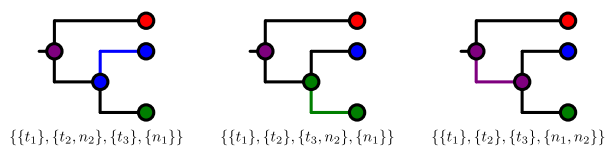

FIG. 2. For the tree in figure 1, the three members of $\mathbf{Q}(\mathcal{T})$ which are not members of $\mathbf{P}(\mathcal{T})$.

$\mathbf{Q}(\mathcal{T}_u)$, unlike $\mathbf{P}(\mathcal{T}_u)$, allows an internal node of $\mathcal{T}_u$ to share its part with no tip of $\mathcal{T}_u$. Suppose $\mathfrak{P}$ is a partition of $N(\mathcal{T})$ and there exists $S \in \mathfrak{P}$ such that $N(\mathcal{T}_u) \cap S$ is nonempty and contains no tip of $\mathcal{T}_u$. Then:

1. $u \in N(\mathcal{T}_u) \cap S$ because if it were not then the S would not obey the connectedness requirement for being a part of a partition of $N(\mathcal{T})$. This is because, if $v \in S \cap N(\mathcal{T}_u)$ and t is the tip of $\mathcal{T}$ in S, then the path from v to t must intersect u.

2. $N(\mathcal{T}_u) \cap S$ is the only member of the set $\{N(\mathcal{T}_u) \cap R : R \in \mathfrak{P}\}$ that contains no tips of $\mathcal{T}_u$, because u can belong to only one member of a partition of $\mathcal{T}_u$.

It follows that $\mathbf{Q}(\mathcal{T}_u)$ is the set of partitions of $\mathcal{T}$ which obey the rules for an admissible partition except that they also allow (but do not insist on) an extra part (whose elements still induce a connected subgraph of $\mathcal{T}_u$) containing $\mathcal{T}$'s root. There is now no need to insist that $\mathbf{Q}(\mathcal{T})$ only be defined if $\mathcal{T}$ is a subtree of some larger tree; it is defined for any tree. Figure 2 shows an example of the extra elements of $\mathbf{Q}(\mathcal{T})$ which are not already elements of $\mathbf{P}(\mathcal{T})$ (and hence already displayed in figure 1).

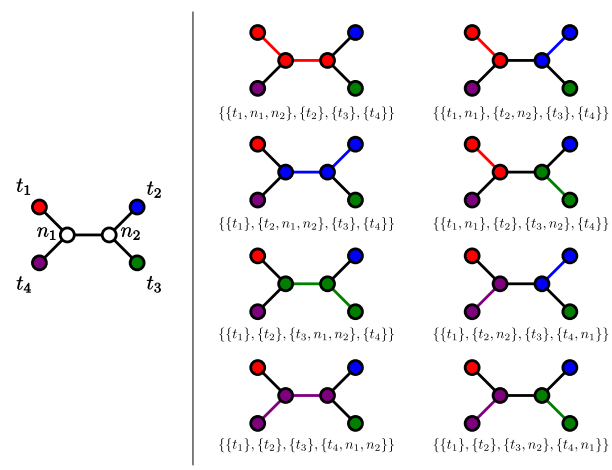

FIG. 3. An unrooted phylogeny (left) and the eight partitions of its node set (right). Coloured branches connect members of the same part.

We will not need to use the definition of $\mathbf{Q}(\mathcal{T}_u)$ again, because it is in obvious correspondence with $\mathbf{P}(\mathcal{T}^*)$. (Recall that $\mathcal{T}^*$ is obtained from $\mathcal{T}$ by attaching a single tip to $\mathcal{T}$'s root.) Compare figure 3 with the full set of partitions displayed in figures 1 and 2 as an illustration of this.

If n is at least 2, then $\mathcal{T}$ has a left subtree $\mathcal{T}_{rL}$ rooted at the left child rL of its root node r and a right subtree $\mathcal{T}_{rR}$ rooted at the right child rR.

Proposition 1. *If $\mathcal{T}$ has at least two tips, then:*

$$|\mathbf{P}(\mathcal{T})| = (|\mathbf{P}(\mathcal{T}_{rL})| \times |\mathbf{P}(\mathcal{T}^*_{rR})|) + (|\mathbf{P}(\mathcal{T}_{rR})| \times |\mathbf{P}(\mathcal{T}^*_{rL})|)$$

Proof. Since $\mathcal{T}$ is not the tree with one node, its root r is internal. First we count the number of partitions where r is in the same part as a tip of $\mathcal{T}_{rL}$. In this case, rL is in that same part by the connectedness requirement for parts: if it were not then the path from that tip to r would go through a node in a different part. The connectedness requirement then also insists that no node of $\mathcal{T}_{rL}$ is in the same part as a tip of $\mathcal{T}_{rR}$, and the number of ways of partitioning the

nodes of $\mathcal{T}_{rL}$ as a subtree of $\mathcal{T}$ such that this is true is just $|\mathbf{P}(\mathcal{T}_{rL})|$. For each of those ways, the number of ways of partitioning the nodes of $\mathcal{T}_{rR}$ is $|\mathbf{Q}(\mathcal{T}_{rR})|$, since some nodes of $\mathcal{T}_{rR}$ can be in the same part as r (and hence rL) and if any are then rR is. Since $|\mathbf{Q}(\mathcal{T}_{rR})|=|\mathbf{P}(\mathcal{T}^*_{rR})|$ the number we are looking for is hence $|\mathbf{P}(\mathcal{T}_{rL})|\times|\mathbf{P}(\mathcal{T}^*_{rR})|$.

An identical argument shows that the number of partitions where r is in the same part as a tip of $\mathcal{T}_{rR}$ is $|\mathbf{P}(\mathcal{T}_{rR})|\times|\mathbf{P}(\mathcal{T}^*_{rL})|$, so the total number is $(|\mathbf{P}(\mathcal{T}_{rL})|\times|\mathbf{P}(\mathcal{T}^*_{rR})|)+(|\mathbf{P}(\mathcal{T}_{rR})|\times|\mathbf{P}(\mathcal{T}^*_{rL})|)$. □

Proposition 2. *If $\mathcal{T}$ has at least two tips, then:*

$$|\mathbf{P}(\mathcal{T}^*)|=|\mathbf{P}(\mathcal{T})|+(|\mathbf{P}(\mathcal{T}^*_{rL})|\times|\mathbf{P}(\mathcal{T}^*_{rR})|)$$

Proof. The root r of $\mathcal{T}$ is an internal node of $\mathcal{T}^*$ connected to a new tip, t (by the construction of $\mathcal{T}^*$). In some partitions of $N(\mathcal{T}^*)$, t is the only member of its part; there are obviously $|\mathbf{P}(\mathcal{T})|$ of these, because counting them is the same problem as counting partitions of the tree rooted at r with t excised.

If t is not the solo member of its part, r is in the same part by the connectedness requirement. The number of ways of partitioning $\mathcal{T}_{rL}$ and $\mathcal{T}_{rR}$ as subtrees of $\mathcal{T}$ in this case are $|\mathbf{Q}(\mathcal{T}_{rL})|$ $(=|\mathbf{P}(\mathcal{T}^*_{rL})|)$ and $|\mathbf{Q}(\mathcal{T}_{rR})|$ $(=|\mathbf{P}(\mathcal{T}^*_{rR})|)$ respectively. The total number of partitions for which t is not the only member of its part is the product of these. □

Since $|\mathbf{P}(\mathcal{T})|$ and $|\mathbf{P}(\mathcal{T}^*)|$ are equal to 1 when $\mathcal{T}$ has one tip, $|\mathbf{P}(\mathcal{T})|$ can now be calculated for any $\mathcal{T}$ by doing a post-order tree traversal, as all that is needed to do the calculations at any node can be obtained by doing the same calculations at both of that node's children. See figure 4 for an example.

Proposition 3. *If $\mathcal{L}_n$ is the ladder tree with n tips, then $|\mathbf{P}(\mathcal{L}_n)|$ is F_{2n-1}, the $(2n-1)$th Fibonacci number, and $|\mathbf{P}(\mathcal{L}^*_n)|$ is F_{2n}.*

Proof. $\mathcal{L}_1$ is the tree with one tip, so $|\mathbf{P}(\mathcal{L}_1)|=1=F_1$ and $|\mathbf{P}(\mathcal{L}^*_1)|=1=F_2$, then proceed by induction. For $n>1$, the two subtrees descended from the root of $\mathcal{L}_n$ are $\mathcal{L}_{n-1}$ and $\mathcal{L}_1$. $|\mathbf{P}(\mathcal{L}_n)|=|\mathbf{P}(\mathcal{L}_{n-1})|\times|\mathbf{P}(\mathcal{L}^*_1)|+|\mathbf{P}(\mathcal{L}^*_{n-1})|\times|\mathbf{P}(\mathcal{L}_1)|=F_{2n-3}\times 1+F_{2n-2}\times 1=F_{2n-1}$ and $|\mathbf{P}(\mathcal{L}^*_n)|=|\mathbf{P}(\mathcal{L}_n)|+|\mathbf{P}(\mathcal{L}^*_{n-1})|\times|\mathbf{P}(\mathcal{L}^*_1)|=F_{2n-1}+F_{2n-2}\times 1=F_{2n}$. □

Enumeration of partitions with a known root part

Having demonstrated how to count the set of partitions or transmission trees compatible with a given $\mathcal{T}$, we now turn out attention to the matter of providing a uniform sample from that set. In order to do this, we need to determine what proportion of the $|\mathbf{P}(\mathcal{T})|$ partitions have the root of $\mathcal{T}$ sharing its part with each tip.

If $E(\mathcal{T})$ is the tip set of $\mathcal{T}$, and $C(\mathcal{T})$ the set of children of r, let $a:\mathcal{P}(E(\mathcal{T}))\to\mathcal{P}(C(\mathcal{T}))$ (with $\mathcal{P}(S)$ representing the power set of S) be the function taking a set of tips of $\mathcal{T}$ to the set of children of r which are ancestors of (or equal to) at least one of those tips.

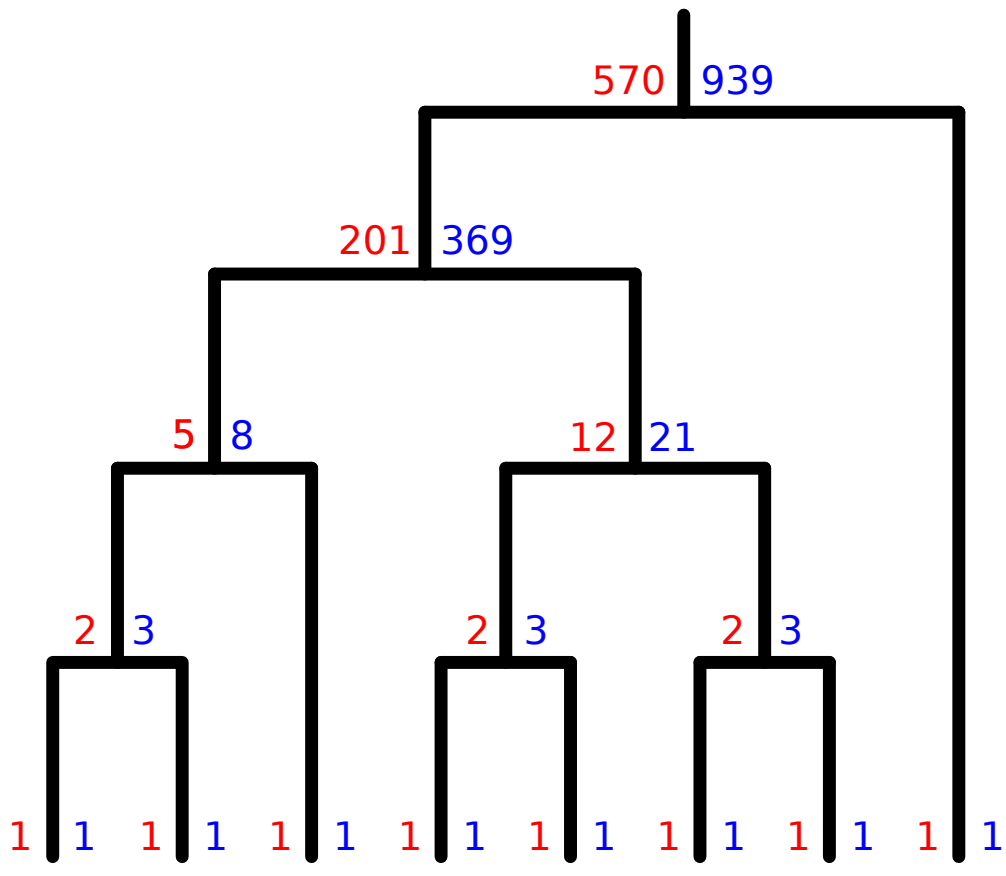

FIG. 4. How to count partitions. At each node u, if $\mathcal{T}_u$ is the subtree rooted at u, then the red number is $|\mathbf{P}(\mathcal{T}_u)|$ and the blue $|\mathbf{P}(\mathcal{T}_u^*)|$. If u is internal and has children uL and uR, $|\mathbf{P}(\mathcal{T}_u)|$ is $(|\mathbf{P}(\mathcal{T}_{uL})|\times|\mathbf{P}(\mathcal{T}_{uR}^*)|)+(|\mathbf{P}(\mathcal{T}_{uR})|\times|\mathbf{P}(\mathcal{T}_{uL}^*)|)$ (the sum of the product of the blue number at uL and the red number at uR, and the product of the blue number at uR and the red number at uL), while $|\mathbf{P}(\mathcal{T}_u^*)|$ is $|\mathbf{P}(\mathcal{T}_u)|+(|\mathbf{P}(\mathcal{T}_{uL}^*)|\times|\mathbf{P}(\mathcal{T}_{uR}^*)|)$ (the sum of the red number at u and the product of the blue numbers at its children).

Let $\{t_1,\ldots,t_n\}$ be the tips of $\mathcal{T}$. For each i let H_i be the set containing just t_i; this may seem redundant but it becomes crucial when relaxing the single sampling assumption as described in the Appendix. If $\mathbf{P}^i(\mathcal{T})\subseteq\mathbf{P}(\mathcal{T})$ is the set of partitions of $N(\mathcal{T})$ that have r in the same part as the membership of H_i, we wish to calculate $|\mathbf{P}^i(\mathcal{T})|$ for all i. Naturally $\sum_{1\le i\le n}|\mathbf{P}^i(\mathcal{T})|=|\mathbf{P}(\mathcal{T})|$. If $\mathcal{T}$ has one tip $t_1\in H_1$, obviously $|\mathbf{P}^1(\mathcal{T})|=1$. For any other $\mathcal{T}$, treating $\mathcal{T}_{rR}$ and $\mathcal{T}_{rL}$ as trees in their own right but whose tips are partially shared with $\mathcal{T}$, we can define $\mathbf{P}^i(\mathcal{T}_{rL})$ (respectively $\mathbf{P}^i(\mathcal{T}_{rR})$) only if $a(H_i)=\{rL\}$ (respectively $a(H_i)=\{rR\}$).

Proposition 4. *Suppose $\mathcal{T}$ has at least two tips. Then:*

$$|\mathbf{P}^i(\mathcal{T})|=\begin{cases}|\mathbf{P}^i(\mathcal{T}_{rL})|\times|\mathbf{P}(\mathcal{T}_{rR}^*)|, & a(H_i)=\{rL\}\\ |\mathbf{P}^i(\mathcal{T}_{rR})|\times|\mathbf{P}(\mathcal{T}_{rL}^*)|, & a(H_i)=\{rR\}\\ 0, & a(H_i)=\emptyset\end{cases}$$

Proof. Suppose $a(H_i)=\{rL\}$. This forces rL to be in the same part as r by the connectedness requirement, and the number of members of $\mathbf{P}(\mathcal{T}_{rL})$ that have this property is $|\mathbf{P}^i(\mathcal{T}_{rL})|$. Such

a choice of partition of $\mathcal{T}_{rL}$ leaves $|\mathbf{Q}(\mathcal{T}_{rR})| = |\mathbf{P}(\mathcal{T}^*_{rR})|$ ways of partitioning the nodes of $\mathcal{T}_{rR}$. The argument if $a(t_i) = rR$ is identical. The final line is included for completeness in the case where $\mathcal{T}$ is a subtree of a larger tree and t_i is a tip of the latter but not the former. □

The value of $|\mathbf{P}^i(\mathcal{T})|$ for all i can be calculated by a similar post-order traversal to that described in the previous section. See figure S1 for an example. Note that with an algorithm to calculate all $|\mathbf{P}^i(\mathcal{T})|$, a separate one to calculate $|\mathbf{P}(\mathcal{T})|$ is not necessary as $|\mathbf{P}(\mathcal{T})|$ is just the sum of the $|\mathbf{P}^i(\mathcal{T})|$.

Sampling uniformly from $\mathbf{P}(\mathcal{T})$

If the post-order traversal above is complete (and its results recorded for all subtrees of $\mathcal{T}$, not merely $\mathcal{T}$ itself), sampling a random partition requires a single *pre*-order traversal. We start with a collection of empty sets $\mathfrak{P} = \{S_1, \ldots, S_n\}$, where each S_i is to contain the set H_i; once the traversal is complete, $\mathfrak{P}$ will be a partition of $N(\mathcal{T})$. The traversal starts at r, and the $|\mathbf{P}^i(\mathcal{T})|$ can be used as a set of probability weights for a draw of the S_i that r belongs to, as they determine, for each i, how many of the $|\mathbf{P}(\mathcal{T})|$ total partitions have r in S_i.

Subsequently, when the traversal reaches another node u with parent uP, and we have already placed uP in S_i, then u must also be placed in S_i if t_i is one of its descendants (by connectedness) or if u is t_i itself. Otherwise, there are $|\mathbf{P}(\mathcal{T}^*_u)|$ ways in which $\mathcal{T}_u$ can be partitioned, since it can be a member of the same part as uP or a member of the same part as each of its tips. $|\mathbf{P}(\mathcal{T}^*_u)| - |\mathbf{P}(\mathcal{T}_u)|$ of these have u in the same part as uP, while the remaining $|\mathbf{P}(\mathcal{T}_u)|$ do not. For each j such that $t_j \in E(\mathcal{T}_u)$, $|\mathbf{P}^j(\mathcal{T}_u)|$ give the numbers of ways in which u can be placed in the same part as t_j. The part for u can then be sampled with probability given by a weight vector that has $|\mathbf{P}^j(\mathcal{T}_u)|$ for each S_j if $t_j \in E(\mathcal{T}_u)$, $|\mathbf{P}(\mathcal{T}^*_u)| - |\mathbf{P}(\mathcal{T}_u)|$ for S_i, and 0 for any other part.

Counting using the reduced Laplacian

Here we give an alternative non-recursive means of calculating $|\mathbf{P}(\mathcal{T})|$ and $|\mathbf{P}(\mathcal{T}^*)|$, using graph theory. Following Levine (2009), let $\tilde{\mathcal{T}}$ be the graph obtained from $\mathcal{T}$ by collapsing all tips to a single node s.

We observe that there is a bijection between $\mathbf{P}(\mathcal{T})$ and the set of spanning trees of $\tilde{\mathcal{T}}$. Suppose $\mathfrak{P} \in \mathbf{P}(\mathcal{T})$ is a partition. The union of the subgraphs induced by each part of $\mathfrak{P}$ is a spanning forest of $\mathcal{T}$ because every node belongs to one part. When $\mathcal{T}$ is transformed to $\tilde{\mathcal{T}}$, this spanning forest becomes a spanning tree: with the tips collapsed to a single node, the resulting graph is connected (because every part of $\mathfrak{P}$ contains one tip) and has no cycles (because every part contains *only* one tip).

Conversely, if $\mathcal{S}$ is a spanning tree of $\tilde{\mathcal{T}}$ then, in "un-collapsing" the node s to recover $\mathcal{T}$, we split $\mathcal{S}$ into n disconnected components. These clearly

form a spanning forest which in turn defines a partition.

Let $u_1,\ldots,u_m$ be the nodes of $\tilde{\mathcal{T}}$, excluding s (although in fact, any single node can be excluded to give the same result), in an arbitrary order. If d_i is the degree of u_i and d_{ij} the number of edges between u_i and u_j (universally 0 or 1 in our case), we now form the reduced Laplacian Δ of $\tilde{\mathcal{T}}$, an $m \times m$ matrix:

$$\Delta_{ij} = \begin{cases} d_i, & i = j \\ -d_{ij}, & i \neq j \end{cases}$$

The matrix-tree theorem then states that the number of spanning trees of $\tilde{\mathcal{T}}$, and hence the cardinality of $\mathbf{P}(\mathcal{T})$, is the determinant of Δ.

The same procedure can be used to count the members of $\mathbf{P}(\mathcal{T}^*)$, with $\tilde{\mathcal{T}}^*$ formed by collapsing the extra tip to s along with the rest. Levine (2009) defines $\tilde{\mathcal{T}}^*$ as the *wired tree* of $\mathcal{T}$, but the extra edge from $\mathcal{T}$'s root to s specified in that definition is optional for our purposes.

As an example, if $\mathcal{T}$ is the tree with three tips and u_1 is its root, then the reduced Laplacian of $\tilde{\mathcal{T}}$ is $\begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}$, which has a determinant of 5, and that of $\tilde{\mathcal{T}}^*$ is $\begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}$ which has a determinant of 8, both as expected (see figures 1 and 3).

This result does not generalise to situations of incomplete or multiple sampling, where parts of a partition can include either zero or more than one tip: in the former case a spanning forest of $\mathcal{T}$, following the transformation to $\tilde{\mathcal{T}}$, remains a forest rather than a tree, and in the latter case it has cycles. Thus, we leave further exploration of this avenue for future work.

Results

Sampling random transmission trees that are consistent with a fixed phylogenetic tree has applications in transmission inference and in phylodynamics. In particular, there has been some work on whether imbalanced phylogenies are indicative of specific kinds of transmission (Colijn and Gardy, 2014; Frost and Volz, 2013; Leventhal *et al.*, 2012; Robinson *et al.*, 2013). It is clear that the phylogenetic tree places some constraints on who may have infected whom, particularly if individuals are treated and become uninfectious at the time of sampling. The current work aids investigations of this nature by permitting quantitative comparison of transmission trees sampled uniformly at random from two different phylogenetic trees.

The shapes of phylogeneties have been related to transmission patterns in a number of studies, as phylogenetic data are an appealing alternative to classical methods, such as contact tracing, to investigate transmission particularly in settings where highly-transmitting individuals may be difficult to identify directly, for example in sexually transmitted or blood-borne infections (Leventhal *et al.*, 2012). In particular, how so-called "superspreaders" (individuals transmitting an infection to a large number of secondary

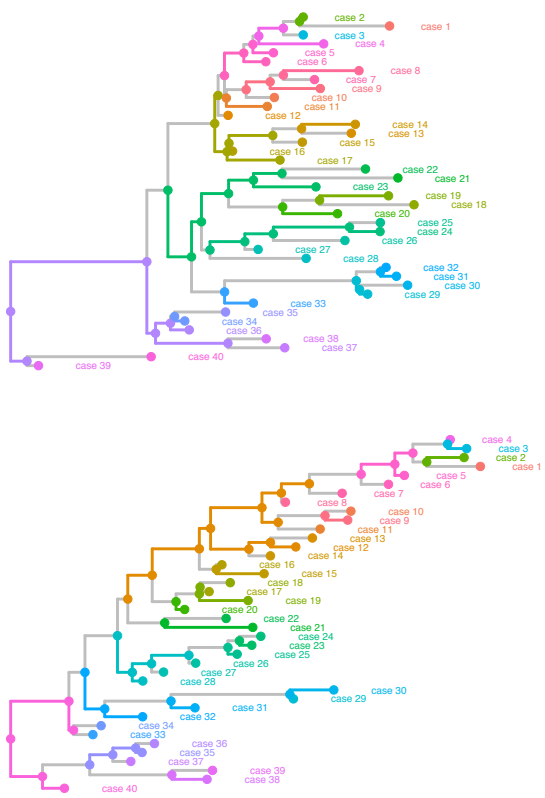

FIG. 5. Yule (top) and biased (bottom) phylogenetic trees with randomly sampled partitions. Each colour corresponds to a part of each partition. Grey edges separate nodes that are in different parts of the partition.

cases), or contact number heterogeneity more broadly, may leave a signature in phylogenetic trees is one important phylodynamic application, particularly in HIV. Several studies have related contact number heterogeneity to the imbalance and cluster patterns in phylogenetic trees, with conclusions that differ depending on assumptions about the network structure and dynamics and the simulation approach (Colijn and Gardy, 2014; Frost and Volz, 2013; Leventhal *et al.*, 2012; Robinson *et al.*, 2013). One of the most commonly used ways to describe the shapes of phylogenetic trees is with their overall asymmetry (imbalance), via, for example, the Sackin index (Sackin, 1972). Indeed, in the phylodynamic literature, this and the number of cherries in the phylogeny have been the primary measures of tree shape. We explored whether there is a systematic difference in the offspring distribution in randomly sampled transmission trees resulting from their asymmetry.

We began with two input phylogenetic trees each with 40 tips, one from a Yule model (a pure branching process) and one from a so-called "biased" model with a bias parameter 0.9. The biased model is a growing tree model; the children of a lineage with a speciation rate r have rates pr and $(1-p)r$. This produces imbalanced trees. The two input trees, along with a randomly sampled partition assuming full sampling and only one tip per host, are shown in Figure 5.

We sampled 300 transmission trees uniformly at random on our two random trees, with full sampling and one tip per host, and compare the distribution of the number of secondary cases infected by a host (offspring distribution). With full sampling, the mean number of secondary cases in a tree is just under 1, because each individual except the source has a single infector.

We find that the relationship between the phylogenetic tree and the dispersion of the offspring distribution depends on whether the timings of infection are restricted. When we make no restrictions on the timing, there is sufficient flexibility in who may infect whom that the two trees have very similar offspring distributions. In contrast, if we constrain the heights of nodes in each tip's part of the partition according to an

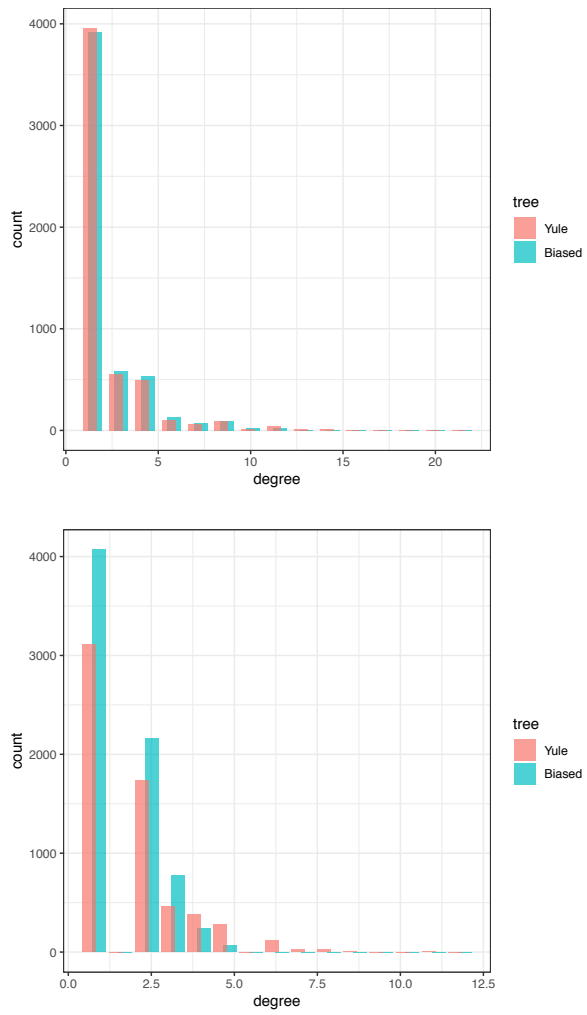

FIG. 6. Offspring distributions from two input phyogenetic trees without (left) and with (right) constraints on the time between infection and sampling such that hosts became non-infectious immediately upon sampling, and had been infectious for a maximum of 3 (Yule) or 3.5 (biased) time units, compared to the mean branch length in these trees of 1.39 and 1.63 time units respectively.

infectious period (using the procedure outlined in the Appendix), the imbalanced tree is more constrained and has a less variable offspring distribution than the Yule tree (Figure 6).

We also compared the transmission trees sampled on the biased and Yule phylogenies directly, using the metric approach developed by Kendall *et al.* (2018). Briefly, the metric is a distance between two transmission trees; the distance is zero if and only if the transmission

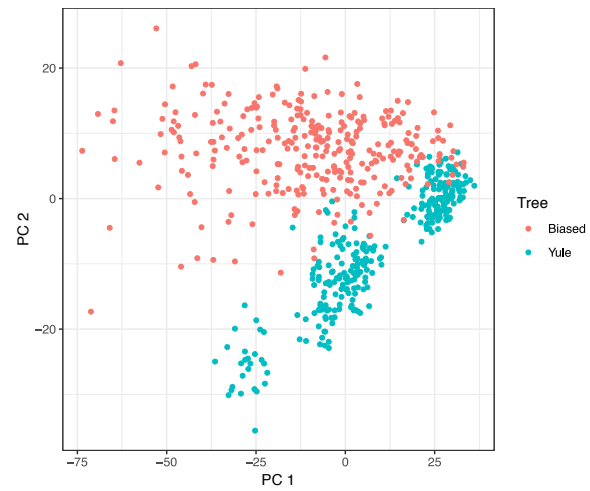

FIG. 7. Principal component analysis plot visualising distances between transmission trees sampled on the Yule and biased phylogenies, without restrictions on the timing of transmission events.

trees are the same (except for some sets of unsampled cases which are not relevant here, as we used full sampling). We compute the distances between all pairs of trees, and visualise the distances using principal component analysis (PCA). Figure 7 shows the results. The Yule and biased phylogenies both admit a "wide spread" of possible transmission trees, but they are strongly separated on the plot. This is a visual illustration of the fact that the structure of the phylogeny places consistent constraints on admissible transmission trees.

We then sampled 500 random phylogenetic trees of 20 tips each and computed the number of transmission trees each one admits. We also computed two common tree shape statistics: the number of cherries and the Sackin imbalance. A cherry is a configuration consisting of two tips and an internal node. Each binary phylogenetic tree has at least one cherry and could have at most

$n/2$ cherries. The Sackin imbalance (Blum and François, 2005; Sackin, 1972) has been defined in several ways, including the total or alternatively the average path length from a tip to the root of the tree. Broadly, the number of possible transmission trees on a fixed phylogeny increases as the Sackin imbalance increases, and declines as the number of cherries increases (cherries are symmetric feature, so trees with higher numbers of cherries tend to have a lower Sackin imbalance). This is, again, under the assumption that there no constraints on the timing of transmission relative to the node's sampling time.

Finally, we compared randomly sampled transmission trees to transmission trees estimated by the *TransPhylo* algorithm (Didelot *et al.*, 2017). We used an outbreak of tuberculosis cases over a 13 year period in Hamburg, Germany, which was previously published (Roetzer *et al.*, 2013) and previously analysed in *TransPhylo* (Didelot *et al.*, 2017). Because the current version of *STraTUS* cannot apply limits on infection times to unsampled cases (see the Appendix), we applied the two algorithms to a 72-tip subtree in which the root node of the epidemic was plausibly infecting a sampled host (see figure S2). (This restriction in *STraTUS* means it will not favour any particular number of unsampled hosts in the transmission tree separating the root node from the first sampled case, regardless of the branch lengths involved. This is very different to *TransPhylo*, so we ensure that the

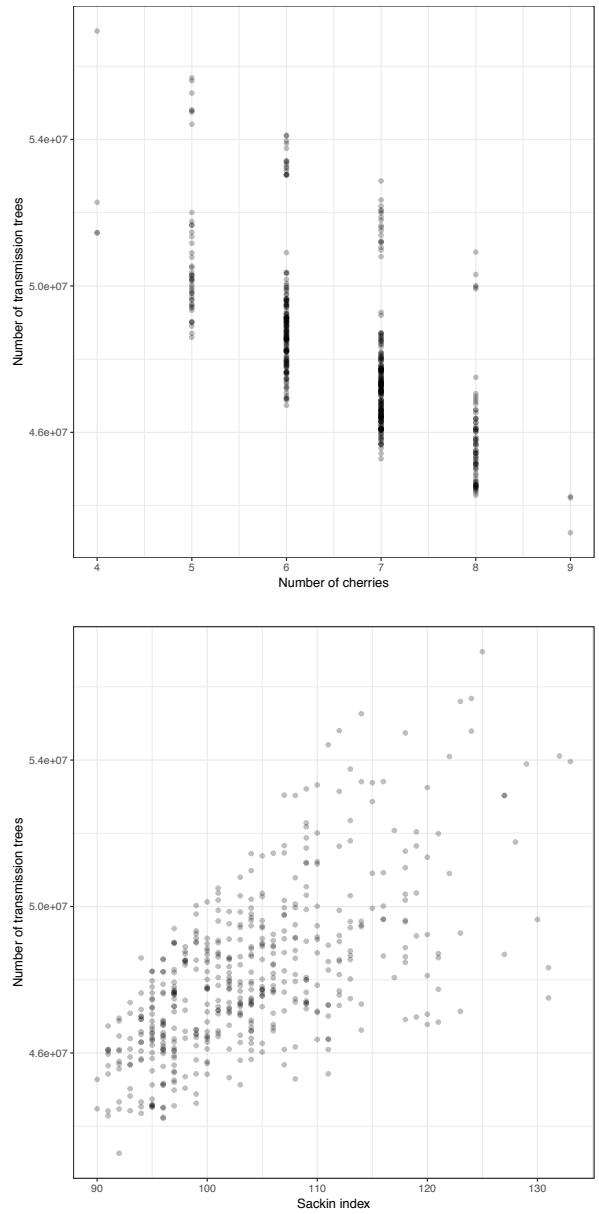

FIG. 8. The number of transmission trees vs the number of cherries (left) and the Sackin measure of imbalance (right) over 500 random phylogenies each with 20 tips.

root case was plausibly sampled in order to make a comparison.) We sampled transmission trees uniformly at random with *STraTUS*, and compared them to the *TransPhylo*-estimated trees. The timed phylogeny was estimated in *BEAST* (Suchard *et al.*, 2018) and was the same as reported in (Didelot *et al.*, 2017), then pruned to the 72-tip subtree. We restricted the maximum possible time between infection and sampling

to 7 years, and assumed that cases become noninfectious upon sampling. We generated multiple *STraTUS* samples for zero, 20, 40 and 60 unsampled hosts, and also with the unsampled count drawn from the empirical distribution of unsampled hosts from *TransPhylo*. The median number of unsampled hosts from *TransPhylo* was 39.

We used the metric and MDS approach outlined above to compare the sets of transmission trees. Figure 9 illustrates the results in 2-dimensional MDS. For 40 unsampled hosts and when the unsampled count was drawn from the *TransPhylo* empirical distribution, the *STraTUS* sample occupies much of the same space as *TransPhylo*, but that the *STraTUS* transmission trees are much more widely distributed. This is not surprising, as the sampling of a *TransPhylo* tree is determined by its posterior probability under a phylodynamic model, whereas *STraTUS* is a cruder, uniform sample from the space of all admissible phylogenies. The *STraTUS* sample with no unsampled hosts, on the other hand, forms a largely distinct cluster in the plot from the *TransPhylo* trees.

We also determined the tree within each group that is closest to the centre of the trees (the geometric median tree; (Jombart *et al.*, 2017)). These are marked on figure 9. Notably, the *STraTUS* sample whose median is closest to the *TransPhylo* median is the one where the unsampled host count was drawn from the *TransPhylo* empirical distribution. These results suggests that it may be possible to use *TransPhylo* to quickly produce an approximate sample of possible transmission trees for a given phylogeny, but that unbiased estimation of the number of unsampled individuals would be necessary.

For the *TransPhylo* and empirical *STraTUS* samples, we show the geometric median transmission trees in Figure 10. While these are different, they share a number of transmission events and features. The distribution of unsampled cases is notably different because *STraTUS* does not take branch lengths into account in placing them while *TransPhylo* does.

Discussion

In this paper, we have explored the mathematics of the set of transmission trees for a known phylogeny, if internal nodes of that phylogeny are not taken to represent infection events, in greater depth and with more rigour than in any previous work. We also give algorithms for uniform sampling of transmission trees. In addition to establishing a firm footing for further theoretical work of this nature, and providing a new method to investigate the relationship between properties of the phylogeny and disease epidemiology, this work has several other potential applications.

The packages *TransPhylo* (Didelot *et al.*, 2014, 2017) and *BEASTLIER* (Hall *et al.*, 2015) both employ MCMC sampling of partitioned trees to estimate transmission trees, for a fixed

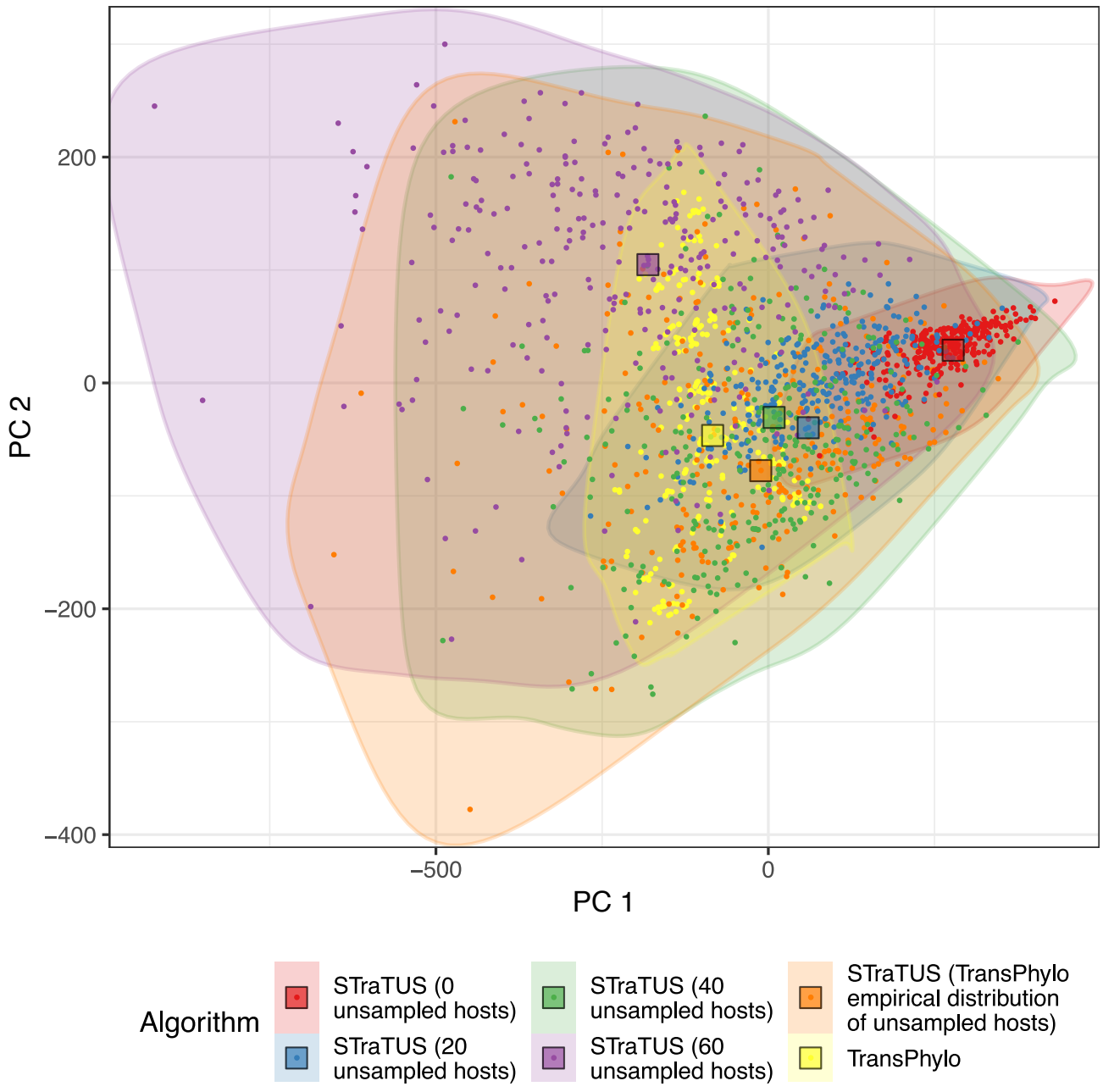

FIG. 9. PCA plot illustrating the distances between transmission trees inferred by *TransPhylo* and sampled using *STraTUS*, derived from the timed phylogenetic tree of the Roetzer outbreak, previously used in (Didelot *et al.*, 2017). The colours indicate the algorithm used and the number of unsampled cases selected in *STraTUS*. The shaded areas represent the extent of the MDS space occupied by the trees in each sample. The squares represent the geometric median tree of each sample.

phylogeny in the former case and a variable one in the latter. The uniform sampling procedure detailed here, perhaps together with recently-developed metrics on phylogeny and transmission tree space (Kendall and Colijn, 2016; Kendall *et al.*, 2018) may prove valuable in the design of improved transition kernels for these algorithms. A uniform sampler for transmission trees may also be useful in a two-stage importance sampling approach of the type employed by Numminen *et al.* (2014), wherein a uniform sample of transmission trees are sampled, given importance

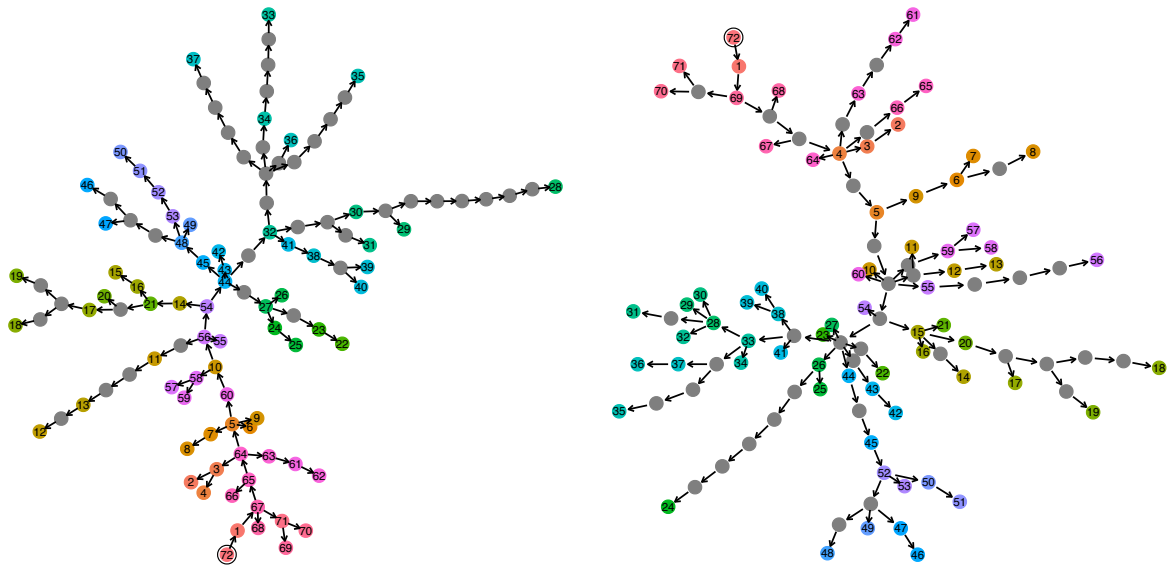

FIG. 10. Geometric median trees from *TransPhylo* (left) and *STraTUS* (right). Grey nodes represent unsampled cases and in each case the index host in the tree is ringed in black. While many individual transmission events differ, there are many points where the differences are "minor" and the trees share small sub-clusters of cases who transmitted to each other in different configurations.

weights according to their likelihood under a model of transmission and then resampled with probability proportional to those weights.

Furthermore, approaches such as *TransPhylo*, *BEASTLIER*, *phybreak* and others make use of a number of models and priors, including for example the natural history of the pathogen (which is used to inform the likelihood based on time between infection and transmission using a generation time), the sampling fraction and sampling process, and a coalescent model for the within-host pathogen evolution. These parameters are difficult to estimate in any single outbreak dataset (particularly in-host evolutionary parameters), and may vary from one outbreak or setting to the next. Re-using past estimates may not solve the problem. The ability to very rapidly sample from all transmission trees consistent with a phylogeny could allow outbreak investigators to rapidly understand which putative transmission events are and are not consistent with genomic data, without making strong assumptions on unknown parameters. When doing this, reasonably accurate quantification of the number of unsampled hosts would appear to be advisible.

The main assumption in transmission tree inference that we are unable to relax is the complete bottleneck at infection. The partition approach basically requires this, as to discard it is to discard the requirement that the region of a phylogeny associated with each host is connected. Without this, any number of transitions amongst the hosts can occur on any branch, and thus the set of transmission trees is infinite. We would argue that that set is actually rather less realistic than the one we present here, as large numbers of reinfection events will be

rare for most pathogens. The importance of the bottleneck assumption in practice has not been extensively studied. Consequential violations of connectedness, where transmission trees exist that are actually impossible under the complete bottleneck assumption (see figure S3), require not just that multiple lineages be transmitted, but that two or more of them are later either transmitted onwards to different hosts, or sampled. How likely this is to happen in practice will vary from pathogen to pathogen and setting to setting; it is more plausible when the "hosts" in the transmission tree are premises rather than individual organisms. It is also unclear whether such an event would ever leave a sufficient signal on the pathogen genome to allow its identification. Nevertheless, methods to estimate transmission trees that do not make the complete bottleneck assumption are available (Worby *et al.*, 2016).

Materials and Methods

Random phylogenetic trees were generated using the `rtree` function in the `R` package *ape* v1.5 and all trees were visualised with *ggtree* v3.7. Transmission trees were compared using the metric of Kendall *et al.* (2018) implemented in *treespace* 1.1.3 (Jombart *et al.*, 2017). Principal component analysis was performed using *ade4* v1.7-13 (Chessel *et al.*, 2004). Sequencing, alignment and BEAST analysis of the *M. tubercolosis* dataset has been previously described (Didelot *et al.*, 2017; Roetzer *et al.*, 2013).

Acknowledgments

This work was supported by European Research Council (advanced grant number) PBDR-339251 and Engineering and Physical Science Research Council (UK) (grant numbers EP/K026003/1 and EPSRC EP/N014529/1).

References

Aldrin, M., Lyngstad, T. M., Kristoffersen, A. B., Storvik, B., Borgan, Ø., and Jansen, P. A. 2011. Modelling the spread of infectious salmon anaemia among salmon farms based on seaway distances between farms and genetic relationships between infectious salmon anaemia virus isolates. *Journal of The Royal Society Interface*, 8(62): 1346–1356.

Blum, M. G. B. and François, O. 2005. On statistical tests of phylogenetic tree imbalance: the sackin and other indices revisited. *Math. Biosci.*, 195(2): 141–153.

Chessel, D., Dufour, A.-B., and Thioulouse, J. 2004. The ade4 package – I: One-table methods. *R News*, 4(1): 5–10.

Colijn, C. and Gardy, J. 2014. Phylogenetic tree shapes resolve disease transmission patterns. *Evol Med Public Health*, 2014(1): 96–108.

Didelot, X., Gardy, J., and Colijn, C. 2014. Bayesian inference of infectious disease transmission from whole genome sequence data. *Molecular Biology and Evolution*, 31(7): 1869–1879.

Didelot, X., Fraser, C., Gardy, J., and Colijn, C. 2017. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular Biology and Evolution*, 34(4): 997–1007.

Frost, S. D. W. and Volz, E. M. 2013. Modelling tree shape and structure in viral phylodynamics. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 368(1614): 20120208.

Giardina, F., Romero-Severson, E. O., Albert, J., Britton, T., and Leitner, T. 2017. Inference of Transmission Network Structure from HIV Phylogenetic Trees. *PLoS*

Comput. Biol., 13(1): e1005316.

Hall, M., Woolhouse, M., and Rambaut, A. 2015. Epidemic reconstruction in a phylogenetics framework: transmission trees as partitions of the node set. *PLOS Computational Biology*, 11(12): e1004613.

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 Computational Biology*, 10(1): e1003457.

Jombart, T., Kendall, M., Almagro-Garcia, J., and Colijn, C. 2017. treespace: Statistical exploration of landscapes of phylogenetic trees. *Mol. Ecol. Resour.*, 17(6): 1385–1392.

Kendall, M. and Colijn, C. 2016. Mapping phylogenetic trees to reveal distinct patterns of evolution. *Mol. Biol. Evol.*, page msw124.

Kendall, M., Ayabina, D., Xu, Y., Stimson, J., and Colijn, C. 2018. Estimating transmission from genetic and epidemiological data: A metric to compare transmission trees. *Stat. Sci.*, 33(1): 70–85.

Klinkenberg, D., Backer, J. A., Didelot, X., Colijn, C., and Wallinga, J. 2017. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLOS Computational Biology*, 13(5): e1005495.

Lau, M. S. Y., Marion, G., Streftaris, G., and Gibson, G. 2015. A systematic bayesian integration of epidemiological and genetic data. *PLOS Computational Biology*, 11(11): e1004633.

Leventhal, G. E., Kouyos, R., Stadler, T., Von Wyl, V., and others 2012. Inferring epidemic contact structure from phylogenetic trees. *PLoS Comput. Biol.*

Levine, L. 2009. The sandpile group of a tree. 30(4): 1026–1035.

Mollentze, N., Nel, L. H., Townsend, S., Roux, K. l., 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. *Proceedings of the Royal Society B: Biological Sciences*, 281(1782): 20133251.

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. 8(11): e1002768.

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. B*, 281(1794): 20141324.

Robinson, K., Fyson, N., Cohen, T., Fraser, C., and Colijn, C. 2013. How the dynamics and structure of sexual contact networks shape pathogen phylogenies. *PLoS Comput. Biol.*, 9(6): e1003105.

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(2): e1001387.

Sackin, M. J. 1972. "Good" and "Bad" Phenograms. *Syst. Zool.*, 21(2): 225–226.

Skums, P., Zelikovsky, A., Singh, R., Gussler, W., Dimitrova, Z., Knyazev, S., Mandric, I., Ramachandran, S., Campo, D., Jha, D., Bunimovich, L., Costenbader, E., Sexton, C., O'Connor, S., Xia, G.-l., and Khudyakov, Y. 2017. QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data. *Bioinformatics*.

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol*, 4(1).

Worby, C. J., O'Neill, P. D., Kypraios, T., Robotham, J. V., Angelis, D. 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(1): 395–417.

Ypma, R. J. F., Ballegooijen, W. M. v., and Wallinga, J. 2013. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics*, 195(3): 1055–1062.