View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Apollo

The Annals of Applied Statistics
2016, Vol. 10, No. 1, 395–417
DOI: 10.1214/15-AOAS898
© Institute of Mathematical Statistics, 2016

RECONSTRUCTING TRANSMISSION TREES FOR COMMUNICABLE DISEASES USING DENSELY SAMPLED GENETIC DATA[1]

BY COLIN J. WORBY[*,†,2], PHILIP D. O'NEILL[*], THEODORE KYPRAIOS[*], JULIE V. ROBOTHAM[‡], DANIELA DE ANGELIS[‡,§,3], EDWARD J. P. CARTWRIGHT[¶], SHARON J. PEACOCK[‖,4] AND BEN S. COOPER[**,††,5]

University of Nottingham[*], Harvard TH Chan School of Public Health[†], Public Health England[‡], MRC Biostatistics Unit[§], Ipswich Hospital NHS Trust[¶], University of Cambridge[‖], University of Oxford[**] and Mahidol-Oxford Tropical Medicine Research Unit[††]

Whole genome sequencing of pathogens from multiple hosts in an epidemic offers the potential to investigate who infected whom with unparalleled resolution, potentially yielding important insights into disease dynamics and the impact of control measures. We considered disease outbreaks in a setting with dense genomic sampling, and formulated stochastic epidemic models to investigate person-to-person transmission, based on observed genomic and epidemiological data. We constructed models in which the genetic distance between sampled genotypes depends on the epidemiological relationship between the hosts. A data-augmented Markov chain Monte Carlo algorithm was used to sample over the transmission trees, providing a posterior probability for any given transmission route. We investigated the predictive performance of our methodology using simulated data, demonstrating high sensitivity and specificity, particularly for rapidly mutating pathogens with low transmissibility. We then analyzed data collected during an outbreak of methicillin-resistant *Staphylococcus aureus* in a hospital, identifying probable transmission routes and estimating epidemiological parameters. Our approach overcomes limitations of previous methods, providing a framework with the flexibility to allow for unobserved infection times, multiple indepen-

Received July 2014; revised November 2015.

[1]Supported in part by funding from the European Community [Mastering Hospital Antimicrobial Resistance (MOSAR) network contract LSHP-CT-2007-037941].

[2]Supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number U54GM088558.

[3]Supported by the UK Medical Research Council (Unit Programme number U105260566).

[4]Supported from UKCRC Translational Infection Research Initiative (MRC Grant number G1000803) and Public Health England.

[5]Supported by The Medical Research Council and Department for International Development (Grant number MR/K006924/1). The Mahidol Oxford Tropical Medicine Research Unit is part of the Wellcome Trust Major Overseas Programme in SE Asia (Grant number 106698/Z/14/Z).

Key words and phrases. Bayesian inference, infectious disease, epidemics, outbreak investigation, transmission routes.

dent introductions of the pathogen and within-host genetic diversity, as well as allowing forward simulation.

1. Introduction. A fundamental aim in the analysis of infectious disease epidemics is to identify who infected whom, however, achieving this is challenging, since transmission dynamics are generally unobserved. A probabilistic estimation of the transmission tree based on all available data offers many potential benefits. In particular, this can lead to improved understanding of transmission dynamics, provide a mechanism to quantify factors associated with heightened transmissibility and susceptibility to carriage and infection, and help identify effective interventions to reduce transmission. Pathogen typing can be used to cluster genetically similar isolate samples, which can rule out potential transmission routes. Whole genome sequence (WGS) data offers maximal discriminatory power through the identification of individual point mutations, or single nucleotide polymorphisms (SNPs), potentially leading to more accurate transmission tree reconstructions than hitherto possible. However, the joint analysis of genetic and surveillance data poses several challenges, as the relationship between epidemic and evolutionary dynamics is complex [Ypma, van Ballegooijen and Wallinga (2013)].

To date, genomic data have primarily been used to analyse transmission at a population rather than an individual level. This typically relies on a broad sample of individuals from a large population, with the aim of estimating past population dynamics over a long period of time. Phylogenetic analyses have been used to infer patterns of large-scale geographic spread [Harris et al. (2010)]. Coalescent theory has been used with such data to estimate, among other things, fluctuations in population size and transmission parameters [Pybus et al. (2001), Volz et al. (2009)]. Methods have also been described to estimate transmission parameters by combining sequence data and time series incidence data [Rasmussen, Ratmann and Koelle (2011)].

In contrast, we focus on individual-level transmission, using high-frequency genomic samples from a subpopulation (e.g., hospital, school, jail, farm, community), with the aim of reconstructing transmission routes. Such sampling presents more of a challenge in terms of resources and data collection. However, with falling sequencing costs, gathering genomic data is rapidly becoming a feasible component of outbreak investigations, as demonstrated by recent studies [Gardy et al. (2011), Köser et al. (2012), Snitkin et al. (2012)]. We aim to estimate the transmission tree, a graph representing the spread of a pathogen between individuals, comprising nodes (cases, which may be defined as infected or colonized persons depending on the context) and directed edges (transmission events). Edges may additionally be associated with a transmission time. A transmission tree may be composed of multiple unconnected subtrees, each representing independent chains of transmission. Each transmission chain has an origin, representing a new introduction of the pathogen into the population. While in some situations it may

be reasonable to regard the tree as fully connected (i.e., only one origin exists), more generally, multiple introductions of the pathogen from external sources must be accounted for.

A number of approaches to reconstruct transmission trees for communicable pathogens using densely sampled genomic data have been described in recent years. Many methods have been based around the construction of phylogenetic trees, which describe the inferred evolutionary relationships between pathogen samples, and can be fit to sequence data under various evolutionary models. The phylogenetic tree is a bifurcating structure in which external nodes represent sampled isolates, while internal nodes represent the most recent common ancestor of its descendants. Internal nodes are similarly linked, such that the structure is fully connected. Since phylogenetic trees may be topologically dissimilar to transmission trees [Pybus and Rambaut (2009), Romero-Severson et al. (2014)], interpreting phylogenetic proximity as epidemiological linkage can be misleading. Furthermore, phylogenetic trees are undirected, leaving ambiguity around the direction of transmission even if the transmission tree is topologically identical.

Phylogenetic trees have been used in conjunction with contact tracing data using ad hoc approaches to rule out possible transmission links [Bryant et al. (2013), Gardy et al. (2011)], while other approaches have developed more formal methods to make use of phylogenetic trees to infer transmission trees. For instance, Ypma, van Ballegooijen and Wallinga (2013) developed a method to sample over both the transmission and phylogenetic tree given a set of sequence data, ensuring both structures remain consistent with one another. This approach required the specification of a model to describe within-host pathogen dynamics, which remain poorly understood for the majority of pathogens. Similarly, Numminen et al. (2014) describe an importance sampling approach in which both phylogeny and transmission tree are sampled from proposal distributions. This approach required sequence data to be partitioned into clusters pre-analysis and the topology of the phylogeny to be fixed, but avoided the computational complexity associated with Markov chain Monte Carlo (MCMC) based methods.

Alternatively, a second class of reconstruction methods avoids phylogenetic tree inference, using models in which transmission routes are weighted by a function of observed genetic distance. Simply identifying the source of infection by selecting the host carrying the most genetically similar sampled isolate has been suggested [Jombart et al. (2011)], although this neglects the role of within-host diversity and sampling time, as well as uncertainty surrounding the times of infection. While more sophisticated approaches allow for uncertainty in transmission time and provide a more realistic model for the accumulation of mutations over time, hosts are characterized by a single pathogen genotype [Mollentze et al. (2014), Morelli et al. (2012), Ypma et al. (2012)]. Jombart et al. (2014) describe a Bayesian data-augmentation approach making use of genetic distance data to infer likely transmission events, dates of infections and unobserved cases. The approach assumes known distributions of the generation interval and time from infection to isolate

collection, and does not allow for within-host diversity or explicitly account for imported cases (though multiple unconnected trees can be allowed for). These assumptions mean that, while the approach may be suitable for an acute infection in an outbreak scenario, it is not appropriate for pathogens such as *S. aureus*, where long-term carriage is common, the generation interval is not well defined, and where within-host diversity can be substantial.

Of the above methods, all assume that a single genotype is sampled from each host, with the exception of Numminen et al. (2014). This assumption can lead to poor tree inference in the presence of within-host diversity [Worby et al. (2014)]. Only the approach developed by Mollentze et al. (2014) can identify importations; the remainder of the methods assume the transmission tree is fully connected. Most methods described assume infection times are known with certainty. It is likely to be extremely useful to relax each of these assumptions in most infectious disease settings. Finally, while the importance sampling method by Numminen et al. (2014) can accommodate various transmission models, the remainder consider instead the probability of a transmission tree linking the set of infected individuals, ignoring the probability of susceptible individuals avoiding infection.

Here we describe a generalized approach to transmission tree reconstruction that overcomes these limitations and makes use of both molecular typing information and known exposure data. A key novelty of our approach is that we model the genetic distances between sequences rather than the microevolution of the sequences themselves. This offers a flexible framework in which multiple independent introductions of the pathogen and within-host diversity may be considered, as well as the transmission process itself. This approach avoids the need to make any assumptions about the within-host pathogen population dynamics, which, in general, are poorly understood. Furthermore, our proposed framework allows data to be simulated forward in time, a feature lacking in the majority of existing methods (with reverse time simulation typically required in phylogenetic methods, and only an incomplete set of genetic distances simulated from other approaches), which is of fundamental importance in predictive modelling and model evaluation.

2. Methods. The importance of identifying transmission pathways in hospital epidemiology is one of the major motivations for our work. We therefore describe our approach for this setting and analyse real and simulated hospital epidemic data. Since infection is often asymptomatic in this setting, even with frequent patient screening, epidemics are only partially observed. Furthermore, patients may be admitted to the ward already infected (importations), which requires consideration of multiple disconnected transmission trees. Our approach accounts for these complications. In line with most literature on hospital-associated infections, we subsequently use the term "colonized" to refer to patients who are either symptomatically or asymptomatically infected with the pathogen.

We observe a set of n patients admitted and discharged from a hospital over a study period. For each patient (j, say), we observe the day of admission t_j^a and

discharge t_j^d, the days and results of screening tests (positive or negative for the pathogen) taken during their stay. We denote the set of all screening results by X. We also suppose that some (not necessarily all) of the positive swabs have a corresponding sequenced isolate, that is, we have genetic information related to some of the positive tests. From a total of n_s sequenced isolates, we derive a symmetric pairwise genetic distance matrix $\Psi = (\psi_{a,b})_{a,b \le n_s}$, with the genetic element $\psi_{a,b}$ giving the genetic distance between isolates a and b. If colonized, the day of colonization for patient j is denoted t_j^c, and the source of infection, s_j, is equal to the ID of the patient from whom the pathogen was acquired or equal to zero if the patient was already colonized on admission. These quantities specify the transmission tree, but are typically unobserved. For patients who are never colonized, $t_j^c = s_j = \infty$. We denote the set of colonization times and routes of infection by T. We can write the likelihood of observing genetic and screening data, given model-specific parameters θ, as

$$\pi(X, \Psi|\theta) = \int_T \pi(\Psi|X, T, \theta)\pi(X|T, \theta)\pi(T|\theta)\,\mathrm{d}T. \tag{1}$$

We now describe the distinct components of our model, which govern the transmission dynamics ($\pi(T|\theta)$), the observation of screening data ($\pi(X|T, \theta)$) and the generation of genetic diversity $\pi(\Psi|X, T, \theta)$.

2.1. *Transmission model.* We first define a stochastic model which describes both pathogen transmission and the genetic distances arising between genotypes sampled from any two individuals. Each patient j, $j = 1, \ldots, n$, is admitted to the ward, independently carrying the pathogen with probability p, and has marker variable ϕ_j, equal to 1 if the patient is positive on admission and zero otherwise. We assumed homogeneous mixing, such that each colonized patient has equivalent contact with each susceptible individual. The rate of transmission to a given susceptible patient on day t is then $\beta C(t)$, where $C(t)$ is the number of colonized patients on day t, and β is the transmission rate per colonized individual. We assumed that individuals colonized on day t may transmit the pathogen from day $t + 1$ until their discharge. Working in discrete time using daily intervals, the probability that a given susceptible patient avoids colonization on day t is $\exp(-\beta C(t))$, thus, acquisition occurs with probability $1 - \exp(-\beta C(t))$. Each patient has the same chance of contacting any other patient in this model, and we note that transmission is often indirect, via the hands of healthcare workers (HCWs) [Albrich and Harbarth (2008), Cooper, Medley and Scott (1999), Pittet et al. (2008)]. Given an individual acquires the pathogen on day t, the probability that the source of transmission is a particular positive individual is simply $1/C(t)$, since it is assumed that colonized patients have an equal potential to transmit. More generally, this probability will be the transmission pressure from the potential source divided by the total transmission pressure at time t. The model for transmission dynamics, T,

can then be given as

$$\pi(T|\theta) = p^{\sum_j \phi_j}(1-p)^{n-\sum_j \phi_j} \times \prod_{i=1}^{n}\left(\mathbf{1}_{t_i^c = t_i^a} + \mathbf{1}_{t_i^c \neq t_i^a} \exp\left\{-\sum_{t=t_i^a}^{\min(t_i^c-1, t_i^d)} \beta C(t)\right\}\right) \times \prod_{\substack{j: t_j^c < \infty, \\ \phi_j = 0}} \left(\frac{1-\exp\{-\beta C(t_j^c)\}}{C(t_j^c)}\right), \tag{2}$$

where $\mathbf{1}_x$ is the indicator function, returning 1 if the condition x is true and zero otherwise.

2.2. *Observation model.* During each patient's stay in the hospital, regular screening is carried out to detect carriage of the pathogen. We assume that the test is highly specific, but imperfectly sensitive—that is, false positive results are not possible, but a positive patient is correctly screened positive with probability z (test sensitivity) [Perry et al. (2004)]. Let $\mathrm{TP}(X,T)$, $\mathrm{FN}(X,T)$ and $\mathrm{FP}(X,T)$ be the total number of true positive, false negative and false positive results in the screening data, respectively, given the set of colonization times. The likelihood of observing the screening results, given test sensitivity and transmission times, is

$$\pi(X|T,\theta) = z^{\mathrm{TP}(X,T)}(1-z)^{\mathrm{FN}(X,T)}\mathbf{1}_{\mathrm{FP}(X,T)=0}. \tag{3}$$

2.3. *Genetic distance models.* We defined the genetic distance to be the observed number of SNPs between isolates, though other metrics are possible. The genetic distance between any two isolates is assumed to be drawn from some probability distribution, which in general can depend on any desired features of the two samples in question or the hosts from whom they were sampled, such as their relatedness in terms of transmission. We assume that genetic distances are perfectly observed, and that insertions, deletions and recombinant sections are removed from the genome such that the genetic distance is representative of the accumulation of SNPs.

The true distribution of the observed number of SNPs between two samples is complex and depends on the mutation rate and the time of their most recent common ancestor, which in turn is dependent on the within-host pathogen population dynamics, as well as the effective transmission inoculum size. Since such factors are still poorly understood for most pathogens, we supposed that the distribution could be approximated by either a Poisson or a geometric distribution, dependent on the relationship between the sampled hosts. This relationship could be modeled a number of ways, but here we focus on two particular models, allowing for genetic diversity to be generated through alternative dynamics.

2.3.1. *Transmission diversity model.* The first model, the transmission diversity model, discriminates between individuals in a transmission chain under the assumption that the expected genetic diversity changes predictably as sampled individuals are further apart in the tree. Typically, one would expect that distances will increase along the chain, due to the accumulation of mutations within each host. Each increase in the tree distance between nodes results in the expected genetic distance changing at a rate governed by a parameter k, which we call the transmission diversity factor. Distances between isolates taken from individuals in unrelated transmission chains are assumed to be drawn from a different specified distribution.

We proposed a distribution to describe the genetic distance between two isolates, given the relationship between their carriers in the transmission tree. For isolates x and y, we defined $t(x, y)$ to be the number of links which separate the isolates in the transmission tree, with $t(x, y) = \infty$ if x and y are sampled from separate chains. For two samples taken from the same host, we have $t(x, y) = 0$. Under the transmission diversity model, we used the following geometric distribution: for $d = 0, 1, \ldots,$

$$P(\psi_{x,y} = d) = \begin{cases} \gamma k^{t(x,y)}\big(1 - \gamma k^{t(x,y)}\big)^d, & t(x, y) < \infty, \\ \gamma_G(1 - \gamma_G)^d, & t(x, y) = \infty, \end{cases} \tag{4}$$

where $\gamma k^{t(x,y)} \in [0, 1]$. Here, the parameter γ_G represents genetic diversity between samples belonging to different transmission chains. The parameter γ is the geometric parameter for genetic distances occurring in the same transmission chain, while k denotes the factor by which this parameter is changed upon an additional transmission link between the samples.

The expected genetic distance between samples is then $(1 - \gamma k^z)/\gamma k^z$ for samples separated by z transmission links or $(1 - \gamma_G)/\gamma_G$ for samples belonging to independent chains. The likelihood contribution for the nth observed sequence is then just the product of probabilities for the $n - 1$ genetic distances to previously observed sequences. Under this model, the likelihood of observing the genetic distance matrix Ψ, given the transmission tree structure, is

$$\begin{aligned} \pi(\Psi | X, T, \theta) = \prod_{y=2}^{n_s} \prod_{x=1}^{y} \big(& \mathbf{1}_{t(x,y)<\infty} \gamma k^{t(x,y)} \big(1 - \gamma k^{t(x,y)}\big)^{\psi_{x,y}} \\ & + \mathbf{1}_{t(x,y)=\infty} \gamma_G (1 - \gamma_G)^{\psi_{x,y}} \big). \end{aligned} \tag{5}$$

We note that in regular circumstances, we would expect $k \leq 1$, indicating steady or increasing diversity along a transmission chain. However, we allow for k to take values greater than 1, as this may highlight sampling bias (e.g., hosts with greater within-host diversity sampled more frequently), which would not be revealed with a fixed upper bound of 1.

The true distribution of genetic distances between independent transmission chains is dependent on the population which enters the hospital already colonized. This distribution will depend upon the strain types circulating in the community and may be multimodal, reflecting clusters of similar strains. In the absence of local and regional sampling data which would be necessary to obtain a more suitable approximation, we use the geometric distribution, assuming strains are more likely to be similar than dissimilar. Our second model is designed to avoid the challenge of approximating this distribution.

2.3.2. *Importation structure model*. The second model, the importation structure model, assumes that imported cases are assigned into genetically similar groups. An individual who acquires the pathogen from another person in a given group is assigned the same group. An importation may belong to a previously observed group, despite not being connected in the transmission chain. The distance between each pair of isolates in a particular group follows the same distribution, regardless of the tree distance between the nodes, while we expect that isolates belonging to different groups to be genetically further apart. The number, and composition, of groups is unobserved, so must be inferred. Under the importation structure model, we have, for $d = 0, 1, 2, \ldots,$

$$P(\psi_{x,y} = d) = \begin{cases} \gamma(1-\gamma)^d, & x \text{ and } y \text{ in the same group,} \\ \gamma_G(1-\gamma_G)^d, & \text{otherwise.} \end{cases} \tag{6}$$

Similar to the previous model, the expected genetic distance between samples is then $(1-\gamma)/\gamma$ for samples within the same group or $(1-\gamma_G)/\gamma_G$ for samples belonging to different groups. It is necessary to introduce some additional notation for this model: let g_j be the group to which colonized individual j belongs (equal to zero if not colonized). We estimate an additional parameter, c, which gives the probability that the strain of an imported case belongs to an existing group. Under this model, the likelihood of observing the genetic distance matrix Ψ, given the transmission tree structure and group memberships g, is

$$\pi(\Psi|X, T, g, \theta) = \prod_{y=2}^{n_s} \prod_{x=1}^{y} \left(\mathbf{1}_{g_x = g_y} \gamma(1-\gamma)^{\psi_{x,y}} + \mathbf{1}_{g_x \neq g_y} \gamma_G(1-\gamma_G)^{\psi_{x,y}}\right). \tag{7}$$

Furthermore, the likelihood of observing n_c groupings among the $\sum_j \phi_j$ importations is

$$\pi(g|\theta) = c^{n_c}(1-c)^{\sum_j \phi_j - n_c}. \tag{8}$$

2.4. *Inference methods*. To allow for unobserved transmission dynamics, namely, the set of transmission days and sources $T = \{t^c, s\}$ (and, additionally, the set of group memberships g under the importation structure model), we used a Bayesian framework and employed a data augmented MCMC algorithm [Tanner

and Wong (1987)] to sample over this space. Individuals with no observed positive swabs may also have been colonized, and we allow for this possibility by sampling over this space. A combination of Metropolis–Hastings and Gibbs sampling was used to draw samples from the parameter space θ, consisting of the parameters $\{p, z, \beta, \gamma, \gamma_G, k\}$ for the transmission diversity model and, additionally, c under the importation structure model. This approach is an extension of the analytical frameworks previously used to estimate transmission parameters given unobserved infection days [Kypraios et al. (2010), O'Neill and Roberts (1999), Worby et al. (2013)]. In addition to sampling transmission days, we specify a model for genetic data in this approach, sampling transmission routes to identify the posterior transmission tree.

Transmission trees were sampled by randomly drawing new colonization days and sources, such that every proposed tree had a nonzero likelihood. Full details of the tree sampling methods, acceptance probabilities and MCMC algorithm are provided in the supplementary material [Worby et al. (2016)]. By calculating the proportion of total samples for which particular transmission routes existed, we derived a tree with edges weighted by posterior probability. The R package "bitrugs" (Bayesian Inference of Transmission Routes Using Genome Sequences) contains code to implement the MCMC algorithm and is included in the online supplementary materials.

Except where mentioned, parameters p, z, γ, γ_G, c were assigned Beta$(1, 1)$ prior densities. The parameters β and k were assumed to be exponentially distributed a priori, with rate 10^{-6}.

2.5. *Data*. We first investigated the performance of our models using simulated hospital data, generated under several different scenarios. Code to simulate data is included in an R package, available in the supplementary materials. We assessed tree accuracy by comparing the simulated true and estimated tree, and examining the receiver operating characteristic (ROC) curve [Krzanowski and Hand (2009)], identifying scenarios in which the model performed well and poorly. We compared our estimated trees to the "uninformed" tree,—that is, an estimate of transmission routes excluding genomic data, assigning each potential source an equal weight. The ROC for the uninformed tree is calculated under the assumption that the times of infection are known, an advantage over our estimation method. Calculating the area under the curve (AUC) and comparing this with the uninformed tree can indicate the improvement in accuracy over the naïve structure.

We then applied our methods to methicillin-resistant *S. aureus* (MRSA) carriage and sequence data collected from a special care baby unit in Cambridge, UK, during an outbreak in 2011. These data comprised a full set of patient admission and discharge days, MRSA carriage screening results and sequenced genomes of a subset of positive results. The genomic data have been described previously by Harris et al. (2013), who combined genomic analysis and contact tracing to estimate routes of infection within and outside of the hospital ward.

3. Results.

3.1. *Simulated data*. We simulated several datasets under the two genetic distance models described in order to determine the ability of our estimation approach to recover the transmission tree as well as the parameter values. We simulated 500 patient admissions over 250 days, and varied model parameters to determine their impact on the ability to identify transmission routes (see supplementary material for more details on simulation). We also investigated the accuracy of tree reconstruction when fitting the model to data simulated under the alternative model. For a range of plausible parameter values we were able to recover the transmission tree well, consistently outperforming the uninformed transmission tree. Under both models, larger outbreaks tended to be associated with more uncertainty surrounding the source of infection. Figure 1 shows a simulated hospital outbreak, comprising several unconnected subtrees. Also shown is the uninformed transmission tree, in which edges are placed with equal weight for all potential sources of transmission, and our reconstruction under the transmission diversity model. While most transmission events are successfully recovered, there is uncertainty within the largest transmission chains which contain several nodes, as well as, in some cases, uncertainty as to whether a case was imported or not. For simulations with an increased transmission rate, a higher number of genetically similar new infections were seen in the ward at any given time, increasing tree uncertainty (Figure 2A). The transmission diversity model allows the length of the transmission chain to have an impact on the expected genetic distances between two given iso-

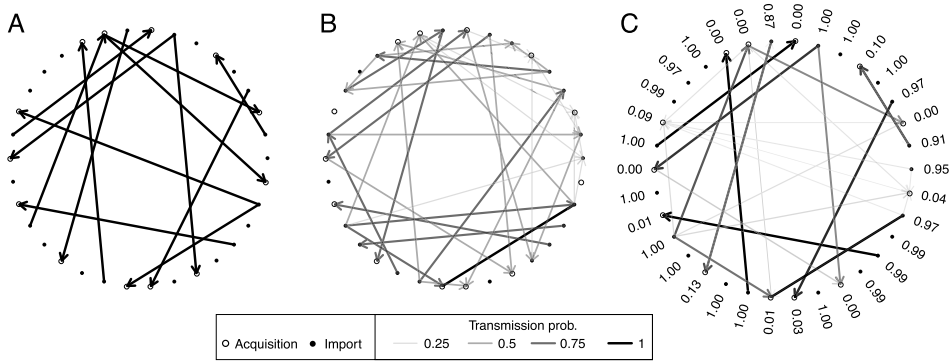

FIG. 1. *A hospital outbreak was simulated, and we attempted to recover the routes of transmission. Patients are represented by open or closed circles, representing acquisitions or importations, respectively, and transmission routes are shown as arrows.* (A) *The true transmission tree.* (B) *The uninformed transmission tree, in which all colonized patients at the time of transmission are considered equally likely sources of infection.* (C) *The estimated transmission tree under the transmission diversity model. Numbers beside each node represent the estimated probability that the individual was positive on admission.*

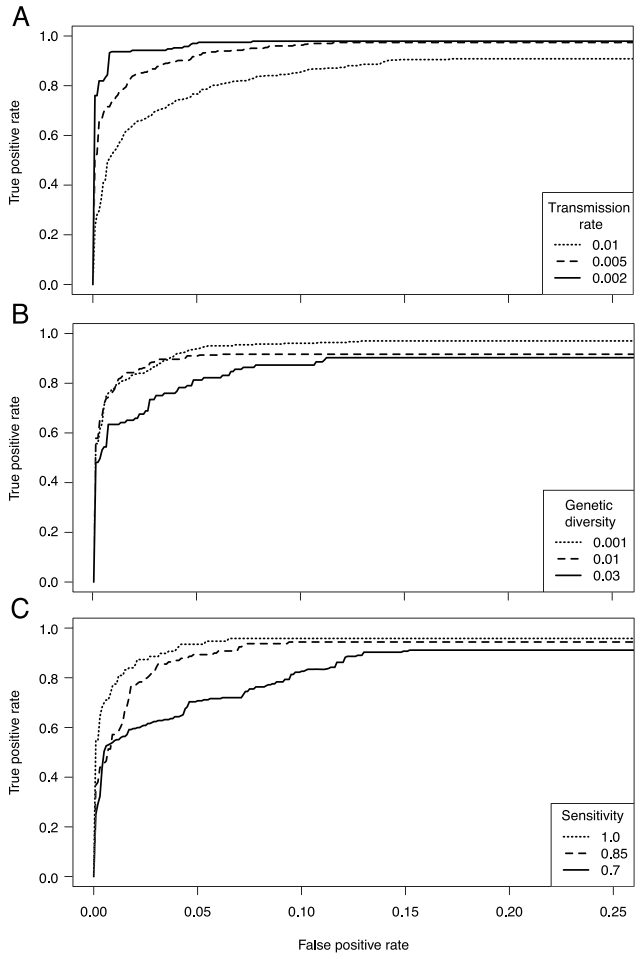

FIG. 2. *ROC curves for estimated transmission trees, based on data simulated under various parameters. We varied transmission rate* (A), *the geometric rate parameter governing between-chain genetic diversity, for which lower values correspond to larger genetic distances* (B), *and test sensitivity* (C). *The ROC curves shown are the average for ten datasets simulated for each scenario.*

lates and therefore allows discrimination between the set of possible sources. For higher transmission rates, transmission chains typically become longer, resulting in the expected genetic distance between isolates approaching the levels expected for unrelated individuals, adding further between-chain uncertainty. Allowing the between-chain expected genetic distance to increase (i.e., reducing γ_G) resulted in improved accuracy (Figure 2B). If imported strains are always highly distinct, then it is straightforward to assign an individual to the correct chain, if not the true source of transmission. Table 1 gives an overview of tree estimation accuracy under various parameter values.

TABLE 1

*Estimated tree accuracy under various scenarios. Each value presented is the mean area under the ROC curve (AUC) for estimated trees under the transmission diversity model, based on 20 datasets simulated under the parameters indicated. Uninformed AUC is based on assigning equal weighting to all available sources. The more accurate reconstruction is highlighted in bold. *Baseline scenario: $p = 0.05$, $z = 0.8$, $\beta = 0.005$, $\gamma = 0.2$, $\gamma_G = 0.05$, $k = 0.8$*

Scenario	Parameters	AUC (uninf.)	AUC (inf.)
Baseline	*	0.67	**0.93**
Low sensitivity	$z = 0.6$	0.67	**0.84**
High sensitivity	$z = 0.9$	0.68	**0.94**
Low transmission	$\beta = 0.001$	0.62	**0.96**
High transmission	$\beta = 0.008$	0.74	**0.91**
Equal diversity ratio	$\gamma = 0.1$, $\gamma_G = 0.1$	0.68	**0.91**
Low diversity ratio	$\gamma = 0.3$, $\gamma_G = 0.1$	0.68	**0.93**
High diversity ratio	$\gamma = 0.3$, $\gamma_G = 0.005$	0.68	**0.96**
No increasing chain diversity	$k = 1$	0.68	**0.93**
Strongly increasing chain diversity	$k = 0.5$	0.69	**0.90**

The importation structure model lends itself to the identification of independent outbreaks rather than individual transmission routes, since, by definition, it may discriminate between groups of similar strains, but assumes a fixed distribution of distances for all samples within a transmission chain. For this reason, tree reconstruction was often more uncertain than under the transmission diversity model, particularly for higher transmission rates. However, the identification of isolate groups was successful for a range of scenarios. In cases with frequent importations, the importation structure model often performed better than the transmission diversity model, particularly when importations were genetically similar to each other. Furthermore, this model generated better tree reconstruction from data simulated under the transmission diversity model than vice-versa. The identification of group membership depended largely on the ratio of within- and between-group expected diversity; the smaller this value, the better the performance (Figure 3).

A key determinant of the transmission diversity model performance was the value of the factor k. The posterior estimate of this parameter was often associated with much uncertainty, especially in the absence of longer transmission chains. Differentiating the exact routes of transmission becomes difficult, or even impossible for values of k close to 1, as genetic similarity along a transmission chain diminishes. Values of k close to zero indicate that a considerable amount of mutation occurs between transmission events, and the genotype within the newly infected individual is very different to that found in the source. We found that tree reconstruction was less successful when k was low (Table 1), and low values of k were typically overestimated.

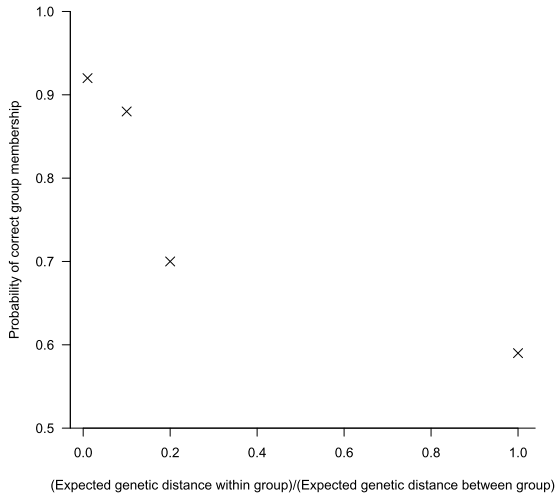

FIG. 3. *Group identification under importation structure model. Data were simulated under a range of within and between group genetic distance distributions, and we estimated the posterior probability that the importation structure model placed an infected individual in the correct group (belonging to the same group as the first importation of that group). Baseline scenario*: $p = 0.05$, $z = 0.8$, $\beta = 0.005$, $c = 0.2$.

In most cases, the ROC curve for estimated transmission trees indicated a considerably better performance than the uninformed tree, demonstrating the gain in information associated with the inclusion of genomic data. However, the tree reconstruction was relatively poor where diversity was defined to be similar for related and unrelated isolates, or when diversity could accumulate quickly in a transmission chain (Table 1). Tree accuracy was relatively poor for lower values of test sensitivity (Figure 2C), but we nevertheless found that our estimates consistently outperformed the uninformed tree (Table 1). However, even with perfect sensitivity, some transmission routes were not recovered, due to colonization and subsequent discharge occurring prior to the next screening time. The degree of uncertainty surrounding even relatively simple trees is notable, reflecting the genetic similarity of linked cases.

We tested sensitivity to our choice of prior distributions by varying the rate parameter of the prior exponential distributions of β and k. We found that neither the parameter estimates nor the estimated transmission tree were affected considerably by varying this value between 10^{-2} and 10^{-10}.

We additionally simulated sequence data under an explicit pathogen evolutionary model. Using the R package "*seedy*" [Worby and Read (2015)], we generated sequence data on top of transmission trees simulated as before. We found that transmission trees could be recovered well, offering a considerable improvement on the uninformed trees (see supplementary material for further details).

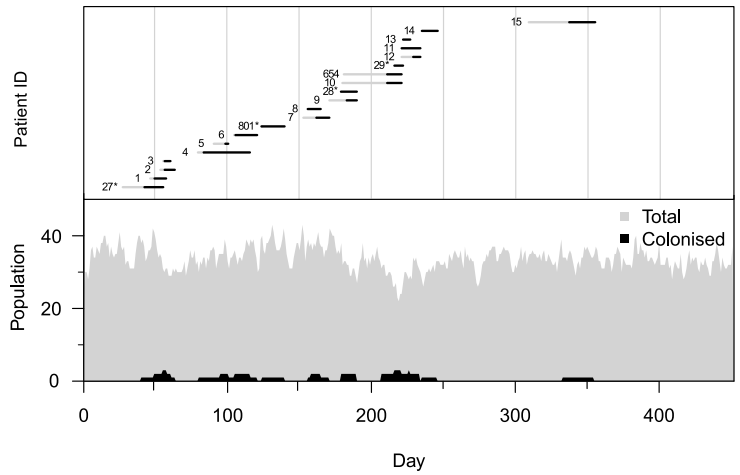

FIG. 4. *Colonized patient episodes in the Rosie hospital neonatal ward. Patients are shown as colonized (black) after their first MRSA positive swab result until the end of their episode. Susceptible patients are shown in grey. Patient marked with an asterisk (*) carry a nonoutbreak sequence type.*

3.2. *MRSA outbreak data from Rosie hospital, Cambridge, UK.* An outbreak of MRSA was observed in 2011 in a special care baby unit at the Rosie Hospital, Cambridge, UK, in which a total of 20 newborn infants were found to be MRSA-positive. We considered a dataset spanning 450 days, including this outbreak, comprising admission and discharge times, as well as MRSA screening results and times, for all patients admitted during this period. A total of 1108 unique patients were admitted to the ward in this period and were swabbed regularly for the presence of MRSA. Figure 4 shows the colonized patient episodes and total population over the study period. Of the 20 patients with positive swabs, 18 had one positive isolate sequenced, and 15 of these were found to be sequence type 2371 (ST2371) (patient numbers 1–15). The remaining three sequenced isolates (carried by patients 27–29) were separated from this outbreak type (and each other) by several thousand SNPs. Two patients (654 and 801) had a positive swab, but no sequenced isolate. During the outbreak investigation, all HCWs were screened voluntarily and with consent, one of whom was found to be MRSA positive. Twenty colonies from this individual were sequenced, revealing carriage of several ST2371 genotypes, differing by up to 10 SNPs (mean pairwise distance 3.9 SNPs). Full details of sequencing and data collection are described in Harris et al. (2013), and sequences were uploaded to the European Nucleotide Archive (www.ebi.ac.uk/ena).

The nonoutbreak sequence types differed by many thousands of SNPs. Fitting the transmission diversity model to these data using a geometric distribution would make the relative likelihood of an observed distance of a much smaller magnitude arising from unrelated transmission chains very low, forcing the model to link all

outbreak strains where possible. This in turn results in an overestimation of the frequency of transmission events. This suggests that a geometric distribution is not an adequate approximation of between transmission chain genetic distances when multiple strain types are present. For this reason, we fitted the transmission diversity model to a restricted dataset, omitting the non-ST2371 strain types. Alternatively, a multimodal distribution could be chosen to account for distant strain clusters, although such a model would likely be overparameterized given the available data. The importation structure model avoids this issue, so we used all available data in this case. For both models, we assumed that test sensitivity was beta distributed with mean 0.8 and standard deviation 0.04 a priori, in line with previous estimates from Worby et al. (2013). We used the sequenced isolates from the colonized HCW to inform our prior density of within-host diversity, γ. All other priors were as described in Section 2.4.

We first ran the MCMC algorithm under the transmission diversity model. Posterior mean estimates and credible intervals of model parameters are summarized in Table 2. We estimated that 1.2% (95% CrI: 0.7%, 1.9%) of patients were positive on admission. The rate of transmission was low, and we estimated a total of 4.8 (3, 7) acquisitions on the ward. Three transmission events had a posterior probability above 0.5, and no transmission was inferred to or from the nonoutbreak types (Figure 5). Around 26% of colonized individuals were the source of one or more secondary cases (Figure 6a). Isolates from patient 654 were not sequenced, therefore we sampled over possible genetic types for this individual. With a high posterior probability (97%), this patient was involved in a transmission event with patient 10, although the direction of transmission was uncertain. We estimated the transmission diversity factor k to be 1.2 (0.7, 1.8), the wide credible interval reflecting the paucity of transmission events, most of which formed a transmission chain of length 1 (Figure 6b). Within-host diversity was estimated to be 3.9 (3.3, 4.6) SNPs, an estimate dominated by the prior density based on the samples from

TABLE 2
Posterior mean estimates and 95% credible intervals for parameters of each model fitted to the Rosie hospital outbreak data

Parameter	**Transmission diversity (95% CrI)**	**Importation structure (95% CrI)**
Probability of importation, p	0.012 (0.007, 0.019)	0.017 (0.009, 0.024)
Test sensitivity, z	0.72 (0.65, 0.79)	0.70 (0.64, 0.77)
Transmission rate $\beta \times 10^{-5}$	89.9 (38.8, 158.2)	80.6 (30.1, 153.7)
Within host/group diversity γ	0.20 (0.18, 0.23)	0.22 (0.19, 0.25)
Between host/group diversity, γ_G	0.17 (0.18, 0.23)	$1.6\ (1.4, 1.9) \times 10^{-4}$
Chain diversity factor, k	1.2 (0.71, 1.82)	—

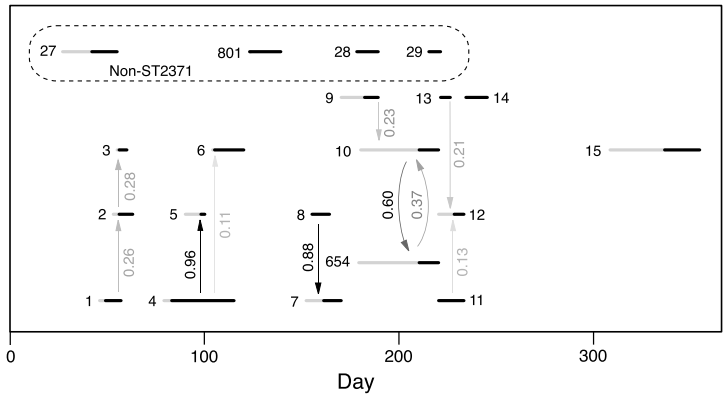

FIG. 5. *Colonized patient episodes in the Rosie hospital neonatal ward. Horizontal bars represent patient episodes, with ID marked alongside. Grey bars denote susceptibility, while black represents the period after the patient's first MRSA positive swab. Arrows denote inferred routes of transmission, with darker arrows representing higher posterior probabilities, the values of which are given alongside. Patients carrying nonoutbreak types are shown at the top of the figure.*

the HCW. As such, the expected distance from source to recipient was approximately 3 SNPs. With the nonoutbreak strain types excluded, the expected distance to unrelated strains was 4.9 (4, 6.1) SNPs. We generated the posterior predictive distributions for the number of observed importations, acquisitions and overall diversity. We found that the true observed values from the dataset fell within the 95% central quantile of the predictive distribution, providing no indication that the model was a poor fit [Worby et al. (2016)].

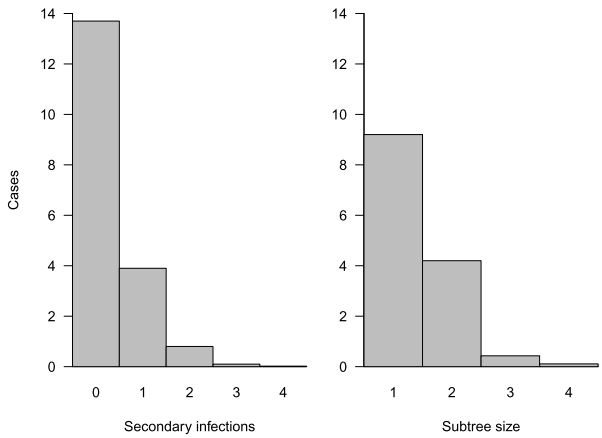

FIG. 6. *Properties of the transmission network, estimated under the transmission diversity model. The posterior distribution of secondary infections for each colonized individual (left) and of the number of connected nodes in each subtree (right).*

The importation structure model placed a high posterior probability on the existence of four groups, reflecting the four sequence types observed in the study. We estimated the expected pairwise distance between isolates belonging to the same group to be 3.7 (3, 4.5) SNPs. Under this model, the probability of importation was estimated to be slightly higher, while the transmission rate was lower. We estimated that patients 1 and 3, who were originally missed by the infection control team at the hospital, were part of the main outbreak group, in accordance with the study by Harris et al. (2013).

4. Discussion. The genetic diversity and structured importation models we have described here allow the combined analysis of genetic and epidemiological data. We applied these methods to the transmission of MRSA in hospitals, demonstrating the simultaneous estimation of model parameters and a transmission tree. More generally, the approaches we have developed can be applied to the analysis of disease transmission in a community where high-frequency sampling of sequence data is available. These methods offer flexibility not available in previous approaches, as they allow multiple introductions of the pathogen into the population, incorporation of within-host genetic diversity, unobserved colonization times, and the provision of estimates of uncertainty for each potential transmission route. While we have used whole genome sequence data, this approach may also be used with lower resolution genetic data, provided a distance metric between isolates can be defined. A major advantage of our framework over existing methods is the ability to simulate forward from our models. This allows one to perform predictive analyses, as well as model evaluation procedures.

A considerable degree of uncertainty was associated with the resulting estimated transmission trees, even for small outbreaks, despite the densely sampled genomic data and well-defined periods of potential contact. As has been previously demonstrated, individual transmission routes are generally unlikely to be identified with high confidence using genetic distance data alone [Worby, Lipsitch and Hanage (2014)]. This reflects the high genetic similarity of individuals in the same transmission chain, and we believe that quantification of uncertainty is of much importance—methods which provide an optimal tree with no measure of uncertainty may be misleading. While we have demonstrated the general improvement in tree accuracy associated with the availability of genomic data, in most cases, much uncertainty is likely to remain regarding transmission routes.

Some previous studies aiming to reconstruct transmission trees using densely-sampled genetic data have used a phylogenetic approach, implicitly assuming that a transmission tree will map closely to the phylogenetic tree [Bryant et al. (2013), Cottam et al. (2008), Gardy et al. (2011)]. However, this assumption may not hold [Pybus and Rambaut (2009), Romero-Severson et al. (2014)]. A fundamental limitation of phylogeny-based approaches is that the relationship between the transmission and phylogenetic trees depends on the within-host evolutionary dynamics

which, in the absence of dense within-host sampling, are not identifiable. By simultaneously sampling over the phylogenetic tree and the transmission tree, one can account for unknown coalescent times and dependencies between genetic distances [Ypma, van Ballegooijen and Wallinga (2013)]. While this approach offers a more realistic model for the emergence of diversity, it also requires a reliable model of within-host pathogen population dynamics. Furthermore, this method requires sampling over the space of phylogenetic trees (and therefore unobserved sequence data), resulting in a considerably more computationally intensive approach than our proposed framework. Even with such a model, the method cannot differentiate between importations and acquisitions, crucial when considering an outbreak in a hospital setting. Data on within-host dynamics are currently scarce, and these dynamics may vary widely between individuals. As such, robust specification of such models is challenging.

Our analysis has some limitations. We have assumed that the source of transmission for each patient must come (indirectly, via HCW) from another patient present on the ward. As Harris et al. (2013) suggested, there is a strong possibility of external sources of transmission in this setting. This would mean that the patient-to-patient transmission rate may be overestimated in our model. Our approach would perform best when all potential contacts are included in the analysis. Additionally, we have used a transmission model that does not allow for heterogeneous rates of transmissibility. We believe that this model is adequate in this setting, and did not affect our primary goal of estimating the transmission tree. We have assumed that clearance of carriage and reinfection are not possible; while it appears unlikely that such events are common in this dataset, incorporating mechanisms for these could be important in other settings and over longer time periods.

Our estimates from the Rosie hospital data suggested that within- and between-host diversity were similar, with the former slightly higher than the latter, suggested by the estimate of $k > 1$. Our estimates of within-host diversity were driven by the HCW, since multiple isolates were not collected from patients. If the HCW was colonized for a long period of time, a higher level of within-host diversity would be expected than within newly colonized infant patients, potentially leading to estimates of $k > 1$. We believe that repeated sampling of each patient would lead to an improved estimate of within-host patient diversity, and that as an estimate of $k > 1$ would be unlikely. We repeated our analysis with k restricted to the interval $[0, 1]$ and found that both parameter estimates and the inferred transmission tree were largely unaffected (supplementary material, Table 2).

We chose simple geometric distributions to represent the genetic diversity both within and between individuals, assuming the probability of each observed sequence was time-homogeneous. We additionally experimented with equivalent Poisson distributions, however, results for the ICU data were very similar using both distributions, although this may not hold for larger datasets with longer transmission chains. While little evidence exists on observed genetic diversity during an epidemic, pairwise genetic distances of the same strain type collected during a

tuberculosis outbreak appear to approximate a geometric distribution [Bryant et al. (2013)], and with a known time to coalescence t and mutation rate μ, the genetic distance should follow a Pois$(2\mu t)$ distribution. With an unknown coalescent time and constant pathogen population size, the genetic distance between contemporaneously sampled genomes should follow a Geometric distribution [Watterson (1975)].

As discussed in Section 2.3.1, the true distribution between independent transmission chains may be multimodal and poorly approximated by a geometric distribution. For this reason, we excluded nonoutbreak sequence types manually before running the transmission diversity model. Our model could be extended in the future to remove this requirement, as for large datasets, and scenarios with concurrent outbreaks belonging to different strain types, this approach would be inappropriate. If local sampling data were available from the community or other regional hospital admissions, we could potentially construct an empirical distribution for the pairwise genetic distances expected between unlinked cases. As yet, such data are typically unavailable, though collection of such data may be feasible in the future.

We have assumed in our analysis that genetic distances are observed without error. In common with all existing tree estimation methods, we assumed that errors arising from sequencing and/or alignment were negligible. In the supplementary material we explored the impact of introducing observation error into the genetic distance matrix, finding that network reconstruction remained largely unaffected by such errors [Worby et al. (2016)].

There are several potential alterations to our model which could be considered and readily incorporated into our framework. The transmission chain diversity model allows the expected genetic distance to increase with number of transmission events and could be reformulated to allow distance to increase linearly or via an alternative relationship. Time between samples could instead be used as the factor by which diversity increases, however, this relationship is complex and only fully understood by accounting for within-host dynamics [Worby, Lipsitch and Hanage (2014)]. Furthermore, since the time between samples from transmission pairs does not vary greatly in this setting, we do not believe it would affect the results significantly. However, in cases where the length of stay (or length of carriage in a nonhospital setting) is long, which would allow times between sample pairs to vary considerably, then such an amendment should be considered.

Furthermore, in creating this model framework, we have assumed that genetic distances are drawn independently, which is not the case in reality. Although in principle this assumption can be relaxed, this would require considerable additional computational complexity. This may be considered in future studies.

Identifying imported cases is challenging, especially when cases are admitted with highly similar strains. In such a setting, our models can exhibit significantly different results—under the importation structure model, an importation of the

same group is more likely than an acquisition soon after admission from another individual on the ward, while under the transmission diversity model, the reverse is true. As such, when strains circulating in the community are very similar to those found in the hospital, the importation structure model will generally perform better, allowing such strains to be clustered importations rather than rapid acquisitions. An intensive care unit admitting patients from elsewhere in the same hospital is an example of a setting where similar strains may be repeatedly imported to the ward. With no prior knowledge of external diversity, it is hard to determine which model is more suitable for identifying importations. However, if both models are run, significant differences between estimated transmission trees suggest that external diversity is similar to that found within the ward. Further data collection would be required to confirm this. The classification of cases as importations or acquisitions is key to the evaluation of infection control procedures, which for healthcare facilities in particular is of great importance. The framework described here can be used to provide evidence towards importation or acquisition in each case using genetic and surveillance data.

Acknowledgements. We are grateful to Simon Harris, Estée Török, Amanda Ogilvy-Stuart, Nick Brown and Cheryl Trundle for assistance with the Rosie Hospital data, as well as the anonymous referee and Associate Editor who provided insightful and constructive comments during the review process. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

SUPPLEMENTARY MATERIAL

Appendix: Transmission tree sampling approach (DOI: 10.1214/15-AOAS898SUPPA; .pdf). Full description of the tree sampling approach, as well as supplementary figures.

Code: Software (DOI: 10.1214/15-AOAS898SUPPSUPPB; .zip). R package "bitrugs" (Bayesian Inference of Transmission Routes Using Genome Sequences), with implementation of data simulation and MCMC algorithm.

REFERENCES

ALBRICH, W. C. and HARBARTH, S. (2008). Health-care workers: Source, vector, or victim of MRSA? *Lancet*, *Infect*. *Dis*. **8** 289–301.

BRYANT, J. M., SCHÜRCH, A. C., VAN DEUTEKOM, H., HARRIS, S. R., DE BEER, J. L., DE JAGER, V., KREMER, K., VAN HIJUM, S. A. F. T., SIEZEN, R. J., BORGDORFF, M., BENTLEY, S. D., PARKHILL, J. and VAN SOOLINGEN, D. (2013). Inferring patient to patient transmission of mycobacterium tuberculosis from whole genome sequencing data. *BMC Infect*. *Dis*. **13** 1–12.

COOPER, B. S., MEDLEY, G. F. and SCOTT, G. M. (1999). Preliminary analysis of the transmission dynamics of nosocomial infections: Stochastic and management effects. *J. Hosp. Infect.* **43** 131–147.

COTTAM, E. M., THÉBAUD, G., WADSWORTH, J., GLOSTER, J., MANSLEY, L., PATON, D. J., KING, D. P. and HAYDON, D. T. (2008). Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proceedings of the Royal Society* (*Series B*) **275** 887–895.

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. *New England Journal of Medicine* **364** 730–739.

HARRIS, S. R., FEIL, E. J., HOLDEN, M. T. G., QUAIL, M. A., NICKERSON, E. K., CHANTRATITA, N., GARDETE, S., TAVARES, A., DAY, N., LINDSAY, J. A., EDGEWORTH, J. D., DE LENCASTRE, H., PARKHILL, J., PEACOCK, S. J. and BENTLEY, S. D. (2010). Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327** 469–474.

HARRIS, S. R., CARTWRIGHT, E. J. P., TÖRÖK, M. E., HOLDEN, M. T. G., BROWN, N. M., OGILVY-STUART, A. L., ELLINGTON, M. J., QUAIL, M. A., BENTLEY, S. D., PARKHILL, J. and PEACOCK, S. J. (2013). Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphyloccus aureus: A descriptive study. *Lancet, Infect. Dis.* **13** 130–136.

JOMBART, T., EGGO, R. M., DODD, P. J. and BALLOUX, F. (2011). Reconstructing disease outbreaks from genetic data: A graph approach. *Heredity* (*Edinb*) **106** 383–390.

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.

KÖSER, C. U., HOLDEN, M. T. G., ELLINGTON, M. J., CARTWRIGHT, E. J. P., BROWN, N. M., OGILVY-STUART, A. L., YANG HSU, L., 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. *New England Journal of Medicine* **366** 2267–2275.

KRZANOWSKI, W. J. and HAND, D. J. (2009). *ROC Curves for Continuous Data. Monographs on Statistics and Applied Probability* **111**. CRC Press, Boca Raton, FL. MR2522628

KYPRAIOS, T., O'NEILL, P. D., HUANG, S. S., RIFAS-SHIMAN, S. L. and COOPER, B. (2010). Assessing the role of undetected colonisation and isolation precautions in reducing methicillin-resistant Staphyloccus aureus transmission in intensive care units. *BMC Infect. Dis.* **10**.

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. *Proceedings of the Royal Society* (*Series B*) **281** 1782.

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, 14. MR3007333

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. *J. R. Soc. Interface* **281** 20141324.

O'NEILL, P. and ROBERTS, G. (1999). Bayesian inference for partially observed stochastic epidemics. *J. Roy. Statist. Soc. Ser. A* **162** 121–129.

PERRY, J. D., DAVIES, A., BUTTERWORTH, L. A., HOPLEY, A. L. J., NICHOLSON, A. and GOULD, F. K. (2004). Development and evaluation of a chromogenic agar medium for methicillin-resistant staphylococcus aureus. *J. Clin. Microbiol.* **42** 4519–4523.

PITTET, D., ALLEGRANZI, B., STORR, J., NEJAD, S. B., DZIEKAN, G., LEOTSAKOS, A. and DONALDSON, L. (2008). Infection control as a major World Health Organization priority for developing countries. *J. Hosp. Infect.* **68** 285–292.

PYBUS, O. G. and RAMBAUT, A. (2009). Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.* **10** 540–550.

PYBUS, O. G., CHARLESTON, M. A., GUPTA, S., RAMBAUT, A., HOLMES, E. C. and HARVEY, P. H. (2001). The epidemic behavior of the hepatitis C virus. *Science* **292** 2323–2325.

RASMUSSEN, D. A., RATMANN, O. and KOELLE, K. (2011). Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comput. Biol.* **7** e1002136, 11. MR2845064

ROMERO-SEVERSON, E., SKAR, H., BULLA, I., ALBERT, J. and LEITNER, T. (2014). Timing and order of transmission events is not directly reflected in a pathogen phylogeny. *Mol. Biol. Evol.* **31** 2472–2482.

SNITKIN, E. S., ZELAZNY, A. M., THOMAS, P. J., STOCK, F., NISC COMPARATIVE SEQUENCING PROGRAM GROUP, HENDERSON, D. K., PALMORE, T. N. and SEGRE, J. A. (2012). Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. *Science Translational Medicine* **4** 148ra116.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–550. With discussion and with a reply by the authors. MR0898357

VOLZ, E. M., POND, S. L. K., WARD, M. J., BROWN, A. J. L. and FROST, S. D. W. (2009). Phylodynamics of infectious disease epidemics. *Genetics* **183** 1421–1430.

WATTERSON, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoret. Population Biology* **7** 256–276. MR0366430

WORBY, C. J., LIPSITCH, M. and HANAGE, W. P. (2014). Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. *PLoS Comput. Biol.* **10** e1003549.

WORBY, C. J. and READ, T. D. (2015). "seedy" (simulation of evolution and epidemiological dynamics): An R package to follow within-host mutation in pathogens. *PLOS One* **10** e0129745.

WORBY, C. J., JEYARATNAM, D., ROBOTHAM, J. V., KYPRAIOS, T., O'NEILL, P. D., ANGELIS, D. D., FRENCH, G. and COOPER, B. S. (2013). Estimating the effectiveness of isolation and decolonization measures in reducing transmission of methicillin-resistant Staphylococcus aureus in hospital general wards. *Am. J. Epidemiol.* **177** 1306–1313.

WORBY, C. J., CHANG, H. H., HANAGE, W. P. and LIPSITCH, M. (2014). The distribution of pairwise genetic distances: A tool for investigating disease transmission. *Genetics* **198** 1395–1404.

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). Supplement to "Reconstructing transmission trees for communicable diseases using densely sampled genetic data." DOI:10.1214/15-AOAS898SUPPA, DOI:10.1214/15-AOAS898SUPPB.

YPMA, R. J. F., VAN BALLEGOOIJEN, W. M. and WALLINGA, J. (2013). Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics* **195** 1055–1062.

YPMA, R. J. F., BATAILLE, A. M. A., STEGEMAN, A., KOCH, G., WALLINGA, J. and VAN BALLEGOOIJEN, W. M. (2012). Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings of the Royal Society* (*Series B*) **279** 444–450.

C. J. WORBY
DEPARTMENT OF MATHEMATICAL SCIENCES
UNIVERSITY OF NOTTINGHAM
NG7 2RD
UNITED KINGDOM
AND
CENTER FOR COMMUNICABLE DISEASE DYNAMICS
DEPARTMENT OF EPIDEMIOLOGY
HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH
677 HUNTINGTON AVE.
BOSTON, MASSACHUSETTS 02115
USA
E-MAIL: cworby@hsph.harvard.edu

P. D. O'NEILL
T. KYPRAIOS
DEPARTMENT OF MATHEMATICAL SCIENCES
UNIVERSITY OF NOTTINGHAM
NG7 2RD
UNITED KINGDOM

J. V. ROBOTHAM
STATISTICS, MODELLING
AND ECONOMICS DEPARTMENT
PUBLIC HEALTH ENGLAND
61 COLINDALE AVENUE
LONDON
NW9 5EQ
UNITED KINGDOM

D. DE ANGELIS
STATISTICS, MODELLING
AND ECONOMICS DEPARTMENT
PUBLIC HEALTH ENGLAND
61 COLINDALE AVENUE
LONDON
NW9 5EQ
UNITED KINGDOM
AND
MRC BIOSTATISTICS UNIT
CAMBRIDGE BIOMEDICAL CAMPUS
FORVIE SITE, ROBINSON WAY
CAMBRIDGE
CB2 0SR
UNITED KINGDOM

E. J. P. CARTWRIGHT
THE IPSWICH HOSPITAL NHS TRUST
HEATH ROAD
IPSWICH
IP4 5PD
UNITED KINGDOM

S. J. PEACOCK
DEPARTMENT OF MEDICINE
UNIVERSITY OF CAMBRIDGE
CAMBRIDGE BIOMEDICAL CAMPUS
CAMBRIDGE
CB2 0QQ
UNITED KINGDOM

B. S. COOPER
CENTRE FOR TROPICAL MEDICINE AND GLOBAL HEALTH
NUFFIELD DEPARTMENT OF MEDICINE RESEARCH BUILDING
UNIVERSITY OF OXFORD
ROOSEVELT DRIVE
OXFORD
OX3 7FZ
UNITED KINGDOM
AND
MAHIDOL-OXFORD TROPICAL MEDICINE RESEARCH UNIT
FACULTY OF TROPICAL MEDICINE
MAHIDOL UNIVERSITY
420/6 RATCHAWITHI RD
RATCHATHEWI DISTRICT
BANGKOK 10400
THAILAND