

1 Genomic infectious disease epidemiology in partially sampled and
2 ongoing outbreaks

3 Xavier Didelot¹, Christophe Fraser^{1,2}, Jennifer Gardy^{3,4}, Caroline Colijn⁵

4 **1** Department of Infectious Disease Epidemiology, Imperial College London, Norfolk Place,
5 London, W2 1PG, United Kingdom

6
7 **2** Oxford Big Data Institute, Li Ka Shing Centre for Health Information and Discovery,
8 Nuffield Department of Medicine, University of Oxford, Oxford OX3 7BN, United Kingdom

9
10 **3** Communicable Disease Prevention and Control Services, British Columbia Centre for
11 Disease Control, Vancouver, British Columbia, Canada

12
13 **4** School of Population and Public Health, University of British Columbia, Vancouver, British
14 Columbia, Canada

15
16 **5** Department of Mathematics, Imperial College, London SW7 2AZ, UK

17 **Abstract**

18 Genomic data is increasingly being used to understand infectious disease epidemiology. Isolates
19 from a given outbreak are sequenced, and the patterns of shared variation are used to infer
20 which isolates within the outbreak are most closely related to each other. Unfortunately, the
21 phylogenetic trees typically used to represent this variation are not directly informative about
22 who infected whom – a phylogenetic tree is not a transmission tree. However, a transmission tree
23 can be inferred from a phylogeny while accounting for within-host genetic diversity by colouring
24 the branches of a phylogeny according to which host those branches were in. Here we extend
25 this approach and show that it can be applied to partially sampled and ongoing outbreaks.
26 This requires computing the correct probability of an observed transmission tree and we herein
27 demonstrate how to do this for a large class of epidemiological models. We also demonstrate how
28 the branch colouring approach can incorporate a variable number of unique colours to represent
29 unsampled intermediates in transmission chains. The resulting algorithm is a reversible jump
30 Monte-Carlo Markov Chain, which we apply to both simulated data and real data from an
31 outbreak of tuberculosis. By accounting for unsampled cases and an outbreak which may
32 not have reached its end, our method is uniquely suited to use in a public health environment
33 during real-time outbreak investigations. We implemented our technique in an R package called
34 TransPhylo, which is freely available from <https://github.com/xavierdidelot/TransPhylo>.

35 Introduction

36 Infectious disease epidemiology is increasingly incorporating genomic data into routine public
37 health practice, using genome sequencing for diagnosis, resistance typing, surveillance, and
38 outbreak reconstruction. In the latter use case, we can draw inferences about the order and
39 direction of transmission based on the presence of mutations common to multiple pathogen
40 isolates (Gilchrist et al., 2015; Croucher and Didelot, 2015). While early works in this area
41 assumed that pathogen genomes from a transmission pair should be identical or near-identical,
42 a number of genomic outbreak investigations revealed the complicating factor of within-host
43 evolution (Ypma et al., 2013; Romero-Severson et al., 2014; Worby et al., 2014).

44 Many important bacterial pathogens have periods of latency, chronic infection, or prolonged
45 asymptomatic carriage, all of which contribute to the generation of within-host genetic diversity
46 (Didelot et al., 2016). *Staphylococcus aureus* is a canonical example, in which single hosts can
47 harbour multiple distinct lineages of the pathogen, each of which may be transmitted onwards
48 (Young et al., 2012; Golubchik et al., 2013; Harris et al., 2013; Tong et al., 2015; Paterson et al.,
49 2015; Azarian et al., 2016). In scenarios where a single host harbours substantial diversity, it
50 can be difficult to infer which other hosts they infected – different lineages may have been
51 transmitted at different points during the donor’s infection and the genome sequenced from the
52 donor may only represent a single lineage captured at the time a diagnostic sample was taken
53 and not the complete set of lineages present within that individual. Indeed, simulation studies
54 have shown that if within-host diversity is ignored, incorrect inferences can be drawn about the
55 transmission events that occurred within an outbreak (Romero-Severson et al., 2014; Worby
56 et al., 2014; Worby and Read, 2015).

57 We have previously introduced a framework for inferring person-to-person transmission events
58 from genomic data that considers within-host genetic diversity (Didelot et al., 2014). We use
59 the genomic data to build a time-labelled phylogeny, which we divided into subtrees, each of
60 which captures the variety of lineages that were present within each host. In other words, the
61 phylogeny is “coloured” with a unique colour for each host, with transmission events represented
62 as changes in colours along a branch. We originally used a simple susceptible-infected-recovered
63 (SIR) model to evaluate the probability of the transmission tree, and we recently showed we
64 can extend our approach to incorporate other types of epidemiological models (Hatherell et al.,
65 2016). A similar approach, developed independently (Hall et al., 2015), couples phylogeny
66 construction and transmission tree inference into a single step.

67 The main limitation of these previous methods is that they assume that all outbreak cases have
68 been sampled and sequenced and that the outbreak has reached its end. These assumptions
69 greatly simplify transmission tree inference, but don’t reflect epidemiological reality. An
70 outbreak is rarely completely sampled – cases may not be reported to public health or they
71 may not have nucleic acid available for sequencing – and genomic epidemiology investigations
72 are frequently unfolding in real-time, meaning an outbreak is being analysed before it is
73 ended. The few methods that can deal with unsampled cases do so at the cost of assuming
74 no within-host diversity (Jombart et al., 2014; Mollentze et al., 2014). Here, we introduce a
75 new Bayesian method for inferring transmission events from a timed phylogeny that can be
76 applied to outbreaks that are partially sampled, ongoing, or both. We solve two problems that
77 arise from these sampling issues: the complexity of calculating the probability of an observed
78 transmission tree under these conditions, and the difficulty in exploring the posterior distribution

79 of possible transmission trees given a phylogeny. Our method also permits the inference of when
80 these transmission events occurred; when coupled with the person-to-person inference, this
81 results in a comprehensive and epidemiologically actionable outbreak reconstruction. Here, we
82 apply our new approach to both simulated datasets and a real-world dataset from the genomic
83 investigation of a tuberculosis outbreak in Hamburg, Germany.

84 **Methods**

85 We use a two-stage approach, first constructing a timed phylogenetic tree \mathcal{P} on the observed
86 sequences and overlaying transmission events (Didelot et al., 2014). Let \mathcal{T} be the transmission
87 tree, \mathcal{P} be the timed phylogenetic tree, θ be the parameters of the transmission and sampling
88 model, and N_{eg} the within-host effective population size.

$$\mathbf{P}(\theta, N_{eg}, \mathcal{T}|\mathcal{P}) \propto \mathbf{P}(\mathcal{P}|N_{eg}, \mathcal{T})\mathbf{P}(\mathcal{T}|\theta)\mathbf{P}(\theta)\mathbf{P}(N_{eg}) \quad (1)$$

89 We compute $\mathbf{P}(\mathcal{P}|N_{eg}, \mathcal{T})$ by separating \mathcal{P} into independent parts, each of which evolves in a
90 different individual host (Didelot et al., 2014; Hall et al., 2015); see below. This separability
91 relies on the assumption of a complete transmission bottleneck, meaning that that within-
92 host genetic diversity is lost at transmission, as is commonly assumed in this context. The
93 central challenge here is therefore to compute $\mathbf{P}(\mathcal{T}|\theta)$ for a general model of transmission:
94 one that allows for both unsampled cases and varying levels of infectivity throughout the
95 course of infection, which is representative of the biological reality for many pathogens. We
96 first illustrate how to do this in a scenario where the outbreak is over; this is a convenient
97 assumption mathematically and makes the derivation simpler. We then proceed to the case
98 where data collection ends at a fixed time before the end of the outbreak, as is the case when
99 analysing an ongoing outbreak.

100 **Basic epidemiological model**

101 The epidemiological process we consider is a stochastic branching process in which each infected
102 individual transmits to secondary cases called offspring (Becker, 1977; Farrington et al., 2003).
103 The number of offspring for any infected individual is drawn from the offspring distribution $\alpha(k)$
104 and we follow previous studies (Lloyd-Smith et al., 2005; Grassly and Fraser, 2008) in assuming
105 that it is a negative binomial distribution with parameters (r, p) . The mean of this distribution
106 is called the reproduction number (Anderson and May, 1992), which is constant and equal to
107 $R = rp/(1 - p)$, and the probability of having k offspring is $\alpha(k) = \binom{k+r-1}{k} p^k (1 - p)^r$. The
108 time span between the primary and any secondary infection is drawn from the generation time
109 distribution $\gamma(\tau)$, where τ is the time since infection of the primary case. The generation time
110 distribution can take any form (Fine, 2003) but a Gamma distribution is often used (Wallinga
111 and Lipsitch, 2007).

112 **Finished outbreak scenario**

113 We first consider the situation where an outbreak follows the model above until there are no
114 more infected individuals; we refer to this as a finished outbreak and we use the star subscript

115 (*) to denote the mathematical quantities associated with this scenario. In this situation, all
 116 individuals are sampled with the same probability π , in which case the time span from infection
 117 until sampling has distribution $\sigma(\tau)$. We want to calculate the probability of a transmission
 118 tree $p_*(\mathcal{T})$. This requires some preliminary quantities.

119 We say that an infected individual is “included” if they are part of the transmission tree by
 120 being either sampled or by leading through transmission to at least one sampled descendant.
 121 Otherwise, we say that an infected individual is “excluded”. Let ω_* be the probability of being
 122 excluded. This means the individual and all of their descendants are unsampled. Considering
 123 the number of offspring k , we have that:

$$\omega_* = (1 - \pi) \sum_{k=0}^{\infty} \alpha(k) \omega_*^k = (1 - \pi) G(\omega_*) \quad (2)$$

124 where $G(z)$ is the probability generating function of the offspring distribution. We model this
 125 as a negative binomial distribution so that $G(z) = \left(\frac{1-p}{1-pz}\right)^r$, but our approach could use other
 126 distributions. We choose the negative binomial distribution to allow individuals to have different
 127 rates at which they are in contact with others (gamma-distributed) combined with a Poisson
 128 distribution of secondary infections given their individual rate. The solution ω_* to Equation 2
 129 is calculated numerically (Supplementary Material).

130 The probability that an individual has d offspring who are included in the process is

$$\mathbf{P}(d \text{ offspring included}) = \sum_{k=d}^{\infty} \binom{k}{d} \alpha(k) \omega_*^{k-d} (1 - \omega_*)^d \quad (3)$$

131 In our final product for $\mathbf{P}(\mathcal{T}|\theta)$, arrived at by induction, each included case will have its own
 132 term. For notational simplicity we define the “modified offspring function” to collect the other
 133 parts of this expression:

$$\alpha_*(d) = \sum_{k=d}^{\infty} \binom{k}{d} \alpha(k) \omega_*^{k-d} \quad (4)$$

134 A good approximation is obtained by taking the sum up to a large value (Supplementary
 135 Material). Note that if $\pi = 1$ then $\omega_* = 0$ and $\alpha_*(d) = \alpha(d)$.

136 We now consider a transmission tree \mathcal{T} generated from the model, which is made of n nodes
 137 corresponding to the included infected individuals (either sampled or unsampled). They are
 138 indexed by $i = 1, \dots, n$. Let $s_i = 0$ if i is unsampled and $s_i = 1$ if i is sampled, in which case
 139 its sampling time is t_i^{sam} . Let t_i^{inf} denote the time when i became infected and d_i denote its
 140 number of included offspring who are indexed by $j = 1..d_i$. The probability of \mathcal{T} given the
 141 parameters θ can be obtained by considering the root ρ of the tree, which has d_ρ offspring, and
 142 the subtrees $\{\mathcal{T}_j\}_{j=1..d_\rho}$ corresponding to each offspring. A recursive form of the probability of
 143 the transmission tree can then be written as:

$$p_*(\mathcal{T}|\theta) = (1 - \pi)^{1-s_\rho} (\pi \sigma(t_\rho^{\text{sam}} - t_\rho^{\text{inf}}))^{s_\rho} \sum_{k=d_\rho}^{\infty} \left(\binom{k}{d_\rho} \alpha(k) \omega_*^{k-d_\rho} \prod_{j=1}^{d_\rho} [p_*(\mathcal{T}_j) \gamma(t_j^{\text{inf}} - t_i^{\text{inf}})] \right). \quad (5)$$

144 The parameters θ appear in the offspring distribution α , the generation time density γ and the
 145 sampling time density σ . The terms in the square brackets do not depend on k , so that we can

146 rearrange the equation using the modified offspring function α_* defined in Equation 4:

$$p_*(\mathcal{T}|\theta) = (1 - \pi)^{1-s_\rho} (\pi \sigma(t_\rho^{\text{sam}} - t_\rho^{\text{inf}}))^{s_\rho} \alpha_*(d_\rho) \prod_{j=1}^{d_\rho} \left[p_*(\mathcal{T}_j) \gamma(t_j^{\text{inf}} - t_i^{\text{inf}}) \right] \quad (6)$$

147 Finally by induction we obtain the probability of \mathcal{T} as a product over all nodes of the
148 transmission tree:

$$p_*(\mathcal{T}|\theta) = \prod_{i=1}^n \left[(1 - \pi)^{1-s_i} (\pi \sigma(t_i^{\text{sam}} - t_i^{\text{inf}}))^{s_i} \alpha_*(d_i) \prod_{j=1}^{d_i} \gamma(t_j^{\text{inf}} - t_i^{\text{inf}}) \right] \quad (7)$$

149 Ongoing outbreak scenario

150 We now consider the situation where an outbreak follows the same model as previously
151 described, until some known time T where observation stops. Whereas individuals were
152 previously all sampled with the same probability π , it is now necessary to account for the
153 fact that individuals who became infected soon before T have a lower probability of being
154 sampled. More formally, the probability of sampling for an individual infected at time t is equal
155 to:

$$\pi_t = \pi \int_0^{T-t} \sigma(\tau) d\tau \quad (8)$$

156 Stopping observation at time T also affects the probability of being excluded, with all individuals
157 infected at $t \geq T$ being excluded.

158 For an individual infected at time t , let ω_t be the probability of being excluded. Note that
159 where $t > T$, $\omega_t = 1$. Before that time, ω_t is not constant, but we know that as $t \rightarrow -\infty$, we
160 should have $\omega_t \rightarrow \omega_*$. We have that:

$$\omega_t = (1 - \pi_t) \sum_{k=0}^{\infty} \alpha(k) \left[\int_0^{\infty} \gamma(\tau) \omega_{t+\tau} d\tau \right]^k \quad (9)$$

161 Let $\bar{\omega}_t = \int_0^{\infty} \gamma(\tau) \omega_{t+\tau} d\tau$. Using the generating function $G(z)$ of the negative binomial
162 distribution of $\alpha(k)$ we have $\omega_t = (1 - \pi_t) G(\bar{\omega}_t)$. We approximate $\bar{\omega}_t$ using a numerical
163 integration (Supplementary Material). Good agreement is found with the expected limit
164 $\omega_{-\infty} = \omega_*$ where ω_* is given in Equation 2.

165 As before, we use the modified offspring function to simplify the notation:

$$\alpha_t(d) = \sum_{k=d}^{\infty} \binom{k}{d} \alpha(k) \bar{\omega}_t^{k-d} \quad (10)$$

166 and obtain a good approximation by taking the sum up to a large value of k (Supplementary
167 Material).

168 With the same recursive reasoning as in the finished outbreak scenario, we have:

$$\mathbf{P}(\mathcal{T}|\theta) = \prod_{i=1}^n \left[(1 - \pi_{t_i^{\text{inf}}})^{1-s_i} (\pi_{t_i^{\text{inf}}} \sigma_t(t_i^{\text{sam}} - t_i^{\text{inf}}))^{s_i} \alpha_{t_i^{\text{inf}}}(d_i) \prod_{j=1}^{d_i} \gamma_t(t_j^{\text{inf}} - t_i^{\text{inf}}) \right] \quad (11)$$

169 where $\sigma_t(\tau)$ and $\gamma_t(\tau)$ are respectively equal to $\sigma(\tau)$ and $\gamma(\tau)$ truncated at time $\tau = T - t$.

170 Inference of transmission tree given a phylogeny

171 The models described above generate transmission trees where each node is an infected
172 individual, each terminal node is a sampled infected individual, and links between nodes
173 represent direct transmission events (Figure 1A). Let us now consider that transmission involves
174 the transfer of only a single genomic variant of the pathogen from the donor to recipient (ie
175 a complete transmission bottleneck) and that sampling involves sequencing a single genome,
176 randomly selected from the within-host pathogen population. The ancestry of the sequenced
177 genomes can then be described as a phylogeny which is made of several subtrees, each of
178 which corresponds to the evolution within one of the included hosts and describes the ancestral
179 relationship between the genomes transmitted and/or sampled from that host (Figure 1B). The
180 probability $\mathbf{P}(\mathcal{P}|\mathcal{T}, N_{eg})$ of a pathogen phylogeny \mathcal{P} given a transmission tree \mathcal{T} and within-
181 host effective population size N_{eg} is therefore equal to the product of the subtree likelihoods for
182 all included hosts (Didelot et al., 2014; Hall et al., 2015), which can be calculated for example
183 under the coalescent model with constant population size N_{eg} (Kingman, 1982; Drummond
184 et al., 2002).

185 Having defined both $\mathbf{P}(\mathcal{T}|\theta)$ and $\mathbf{P}(\mathcal{P}|\mathcal{T}, N_{eg})$, we can now perform Bayesian inference of the
186 transmission tree \mathcal{T} given a timed phylogeny \mathcal{P} using the decomposition in Equation 1. Although
187 a timed phylogeny is not directly available, there are powerful approaches readily available to
188 reconstruct it from genomic data (Drummond et al., 2012; Bouckaert et al., 2014; Biek et al.,
189 2015; To et al., 2016). As in our earlier work (Didelot et al., 2014), we can approach this
190 problem by coloring the phylogeny with one color for each host (Figure 1B); however, since
191 we now consider that some hosts may not have been sampled, the number of infected hosts
192 and therefore the number of colors is not known. In other words, the parameter space is not of
193 fixed dimensionality, and exploring it with a Monte-Carlo Markov Chain (MCMC) requires that
194 we include reversible jumps that change the number of hosts in the transmission tree (Green,
195 1995). Our proposal for adding new transmission events is uniformly distributed on the edges
196 of the phylogeny \mathcal{P} . Our proposal for removing transmission events is uniformly distributed on
197 the set of transmission events that can be removed without invalidating the transmission tree.
198 In a transmission tree \mathcal{T} with n hosts and $\sum_{i=1}^n s_i$ sampled hosts, there are $n - \sum_{i=1}^n s_i$ such
199 removable transmission events. The Metropolis-Hastings-Green ratio for the MCMC move from
200 \mathcal{T} to \mathcal{T}' by adding a transmission event is therefore equal to:

$$\alpha_{\mathcal{T} \rightarrow \mathcal{T}'} = \min \left(1, \frac{\mathbf{P}(\mathcal{T}'|\theta) \mathbf{P}(\mathcal{P}|\mathcal{T}', N_{eg})}{\mathbf{P}(\mathcal{T}|\theta) \mathbf{P}(\mathcal{P}|\mathcal{T}, N_{eg})} \frac{|\mathcal{P}|}{n + 1 - \sum_{i=1}^n s_i} \right) \quad (12)$$

201 where $|\mathcal{P}|$ denotes the sum of the branch lengths of the phylogeny \mathcal{P} . Conversely, the acceptance
202 ratio of the MCMC update from \mathcal{T} to \mathcal{T}' by removing a transmission event is:

$$\alpha_{\mathcal{T} \rightarrow \mathcal{T}'} = \min \left(1, \frac{\mathbf{P}(\mathcal{T}'|\theta) \mathbf{P}(\mathcal{P}|\mathcal{T}', N_{eg})}{\mathbf{P}(\mathcal{T}|\theta) \mathbf{P}(\mathcal{P}|\mathcal{T}, N_{eg})} \frac{n - \sum_{i=1}^n s_i}{|\mathcal{P}|} \right). \quad (13)$$

203 Within each MCMC iteration, additional standard Metropolis-Hastings moves are used to
204 estimate the first parameter r of the Negative binomial distribution for the number of offspring
205 (using an Exponential(1) prior), the second parameter p of the Negative binomial distribution
206 of the number of offspring (using a Uniform([0,1]) prior), the probability of sampling π (using a
207 Uniform([0,1]) prior), and the within-host effective population size N_{eg} (using an Exponential(1)
208 prior).

209 Results

210 Example application to a simulated dataset

211 We simulated an outbreak in which the generation time distribution had a gamma distribution
212 with a mean of 1 year, with a negative binomial offspring distribution with parameters
213 ($r = 2, p = 0.5$), such that the reproduction number was $R = 2$. We set the sampling density
214 at $\pi = 0.5$ with a sampling time distribution identical to the generation time distribution.
215 The simulation was stopped after $n = 100$ genomes had been sampled, which happened at
216 time T . The corresponding phylogeny (Figure 2A) was used as input for our transmission tree
217 inference algorithm with the date T used as described in the “ongoing outbreak scenario” in the
218 Methods section. Performing 50,000 MCMC iterations took less than an hour on a standard
219 computer. The mean posterior of the sampling proportion π was 0.48 with a 95% credibility
220 interval of [0.36,0.59]. The mean posterior of the reproduction number R was 2.168367 with a
221 95% credibility interval of [1.75,2.65]. The estimates of these two key parameters of our model
222 are therefore in excellent agreement with the true values used to perform the simulation.

223 Out of the $n = 100$ sampled individuals, only 53 were infected by another sampled individual;
224 for the majority of these links, our algorithm inferred the existence of the link with high
225 posterior probability, with only nine pairs being given a probability < 0.2 and 15 pairs being
226 given a probability < 0.5 (Figure 2B, red curve). Conversely, for the 9847 pairs of sampled
227 individuals for which a link did not exist in the simulated data, most were given a very small
228 probability of a link in the posterior distribution of transmission trees, with only nine pairs
229 being given a probability > 0.5 (Figure 2B, blue curve). If we consider 0.5 as the probability
230 threshold for when transmission was inferred, our method had a specificity (true negative rate)
231 of 99.9% and a sensitivity (true positive rate) of 72%. The area under the receiver operating
232 characteristic (ROC) curve was 98.97%. These results demonstrate that in this specific example
233 our algorithm was able to infer the correct transmission links with high accuracy, in spite of
234 having information about only a proportion $\pi = 0.5$ of infected individuals. It should be noted
235 that this application represents a best case scenario, since the phylogeny is known exactly,
236 whereas for real epidemiological investigations it would need to be inferred from sequences,
237 adding noise and uncertainty.

238 Evaluation of performance using multiple simulated datasets

239 We repeated the simulation described above for values of the sampling density π varying from
240 0.1 to 1 by increments of 0.01, while leaving the reproduction number constant at $R = 2$. For
241 each of the 90 simulated datasets, we applied our algorithm to estimate the values of both π
242 and R (Figure 3). We found that the estimate of R remained fairly constant as it should, while
243 the estimate of π increased as the correct value of π was increased. There was no sign of a
244 bias in the estimates up to $\pi = 0.6$, but higher values of π were consistently underestimated,
245 with the value of R being slightly overestimated in compensation. We attribute this bias to
246 the difficulty in assessing with certainty whether all cases have been sampled in a transmission
247 chain, since there always remains a possibility that an unsampled individual may have acted as
248 intermediate. This small bias also reflects our choice of prior for π , which was uniform between
249 0 and 1, and the fact that only 100 genomes were used in each simulation.

250 We also performed simulations in the converse situation where the sampling density was kept
251 constant at $\pi = 0.5$ but the reproduction number was increased from $R = 1$ to $R = 11$ by
252 increments of 0.1. For each of the 100 simulated datasets, our method was applied and the
253 inferred values of π and R were recorded (Figure 4). Although there was once again a slight
254 bias towards underestimating the sampling density π , its 95% credibility intervals always covered
255 the correct value of $\pi = 0.5$. The inferred values of R were accordingly overall slightly upward
256 biased, although they followed almost linearly the correct values used for simulation. The 95%
257 credibility intervals for R almost always included the correct values. We conclude from these
258 results that our algorithm performs well despite being tested in difficult situations, with only
259 100 sampled genomes, unknown proportions of unsampled cases, uninformative priors, and very
260 large intervals of values being used in the simulations. A small outbreak with high sampling
261 density and a larger outbreak with lower sampling density can often look similar, especially in
262 the first stages of an ongoing outbreak, but our algorithm is able to distinguish between these
263 two scenarios with good accuracy.

264 Application to a *Mycobacterium tuberculosis* outbreak dataset

265 We applied the method to a previously reported tuberculosis outbreak (Roetzer et al., 2013).
266 We used BEAST (Drummond et al., 2012) to infer a timed phylogeny from the published
267 data (Figure S1). In determining the best priors for the densities of the times between
268 becoming infected and infecting others (the generation time) and between becoming infected
269 and becoming known to the health care system (sampling time), we considered both clinical
270 aspects of tuberculosis disease and aspects of the epidemiological investigation. The outbreak
271 lasted 13 years, during which active case finding was used to identify individuals with prior
272 exposure to known cases. An early report on this outbreak (Diel et al., 2004) noted that many
273 cases were identified for reasons not connected to their tuberculosis infection, such as presenting
274 to health care with other symptoms, to obtain a health certificate, or to enter a detox program.
275 We therefore used a Gamma distribution for the sampling time density, with a shape parameter
276 1.1 and rate 0.4. The generation time for tuberculosis should reflect a chance of relatively rapid
277 progression from infection to active disease and hence to the opportunity to infect others, but
278 also a possibility of infection leading to a long latent period before progression (Barry et al.,
279 2009). We therefore used a Gamma function with shape parameter 1.3 and rate parameter 0.3
280 for the generation time density. We ran 100,000 MCMC iterations. The MCMC traces are
281 shown in Figure S2.

282 Figure 5 shows the consensus transmission network for the real-world tuberculosis outbreak
283 (Roetzer et al., 2013) and Figure 6 shows the inferred numbers of unsampled cases along with
284 the reported cases through time. While most cases were sampled, reflecting a robust public
285 health investigation, we estimate that early in the outbreak, several unsampled individuals
286 were contributing to transmission. During this period, the two major clades of the phylogeny
287 diverged. Figure 6A recapitulates the two major waves of the outbreak – an early peak around
288 1998 and a second pulse from 2005 onwards – each with a small portion of inferred unsampled
289 cases. While the number of unsampled individuals was small, the method does allocate key
290 transmission events to unsampled cases, particularly early in the outbreak, suggesting that
291 screening and investigation earlier in the outbreak was not as comprehensive as it eventually
292 became. This is to be expected, as outbreak management efforts typically intensify as the
293 number of cases grows.

294 Figure 6B shows the posterior times between an individual becoming infected and infecting
295 others – the generation time – and the posterior time intervals between infection and sampling
296 – the infectious period, with priors shown in grey. Our observed generation times are variable,
297 which reflects the clinical history of tuberculosis – an infection that can progress rapidly to
298 active, infectious disease or that can have a asymptomatic, non-infectious latent period of
299 variable length. We used a gamma function as a prior, with mode strictly greater than 0, but
300 the posterior generation times have a mode of 0, suggesting a relatively high portion of those
301 who go on to infect others have a rapid progression to from infection to active disease. It
302 is important to note that the posterior generation times are only an indicator of the inferred
303 natural history of tuberculosis *among those with active disease who were sampled*; individuals
304 who were infected but did not progress to active disease and those who never presented to care
305 and were not sampled do not appear in the dataset, and those who did not infect others do not
306 appear in the cases behind the inferred generation times. The mean posterior generation time
307 was 1.0 years with a standard deviation of 1.36 years. The posterior times between becoming
308 infected and becoming known to health authorities also differ from the prior assumption; they
309 have a mean of 1.4 years and standard deviation of 2 years. Sampling times are distinct from
310 the prior but are affected by a change in the prior assumption.

311 Where inferred infectors are sampled cases with associated clinical and/or epidemiological data,
312 an advantage of our approach is that it allows comparison of the relative contributions of
313 different groups of individuals to the burden of transmission. Figure S3 shows the inferred
314 per-case transmission stratified by several characteristics of the cases (Roetzer et al., 2013):
315 individuals' AFB smear status (a measure of how many bacilli are found in their sputum, if
316 any), HIV status, abuse of alcohol or other drugs, and whether the individual had a permanent
317 domestic residence. Our method did not detect significant differences in secondary infections
318 arising from smear-positive and -negative cases, between substance users and non-substance
319 users, and between stably or transiently housed individuals. However, consistent with the fact
320 that HIV-positive patients tend to be less infectious with tuberculosis, we find that HIV-positive
321 individuals transmitted somewhat fewer cases on average than HIV-negative individuals. Due
322 to the small number of HIV-positive cases – only five individuals were HIV-positive in this data
323 – the estimates are much more variable than the estimates for HIV-negative cases. Many more
324 clinical or demographic factors might impact transmissions, such as the presence of cavitary
325 disease and the reported number of social contacts, but these data were unavailable for the
326 present analysis.

327 Results in Figure S3 do not reflect differences in transmission rates given contact with others,
328 because we do not know about exposures that did not result in infection. We also do not have
329 information about behaviours that might modulate transmission. For example, if smear-positive
330 cases sought and obtained treatment more rapidly than smear-negative cases, or were more
331 unwell and had more limited activities, their transmission rate per contact could be higher than
332 their smear-negative counterparts but they might still contribute fewer onward transmissions.
333 The posterior sampling density is $\pi = 0.93$ with a standard deviation of 0.05, consistent with
334 a very densely-sampled outbreak in a high-resourced setting with good case finding. Posterior
335 estimates of π depend somewhat on priors for σ .

336 Discussion

337 We have described a new methodology for reconstructing who infected whom based on genomic
338 data from an infectious disease outbreak. The novelty of this approach, which extends our
339 earlier work in the area, is that it now accounts for both the possibility of some cases not having
340 been sampled and the possibility that more cases may occur in the future. Addressing these
341 issues overcomes key hurdles in using genomic data to reconstruct disease transmission events
342 during a real-time public health response. In these situations, a case may not be sequenced
343 due to a lack of clinical specimen or otherwise sequenceable material, while cases might go
344 unsampled for various reasons, including subclinical, or asymptomatic, infections for which an
345 individual may not seek care or a diagnosis in another jurisdiction. Furthermore, following early
346 proof-of-concept retrospective studies, genomic epidemiology is now being used to prospectively
347 understand outbreaks, as in the recent outbreak of Ebola (Gire et al., 2014). Allowing inference
348 before the end of the outbreak turns our method into a real-time, actionable approach.

349 Our methodology is based on an explicit transmission model which makes a number of
350 assumptions, some of which could be relaxed if required by specific applications. A first example
351 is the fact that in our model the reproduction number R remains constant throughout the
352 outbreak, whereas in many situations the reproduction number varies over time and quantifying
353 these variations is of great epidemiological importance (Cori et al., 2013). This could be
354 incorporated in our method relatively easily, for example assuming stepwise changes or some
355 predetermined parametric function for $R(t)$. A second example concerns the observation of
356 cases, which we assumed to happen with probability $\pi(t)$ for an individual infected at time
357 t with $\pi(t)$ reflecting the impossibility of observing cases happening after the time T when
358 observation stops and the lower probability of observing cases soon before T (Equation 8). It is
359 often difficult in epidemiological studies to know the real function $\pi(t)$, but in situation where
360 for example surveillance did not start before a certain date, the function $\pi(t)$ we used here could
361 be updated to reflect this. There are also a few other assumptions in our model that would be
362 more difficult to relax, such as the complete transmission bottleneck which considers that only
363 a single pathogen variant is transmitted from the donor to the recipient of each transmission
364 event.

365 A key feature of our methodology is that it proceeds in two steps – first, genomics data is used
366 to reconstruct a phylogenetic tree, and second, likely transmission events given the phylogeny
367 are inferred. There are both advantages and disadvantages to this approach, compared to the
368 more theoretically accurate joint inference of phylogenetic and transmission trees (Hall et al.,
369 2015). Our two-step approach makes it difficult to pass the uncertainty in the phylogenetic
370 reconstruction on to the transmission analysis. This is especially relevant if the time-labelled
371 phylogeny is inferred not using a point estimation procedure (Fourment and Holmes, 2014; To
372 et al., 2016), but rather with a Bayesian sampling method (Drummond et al., 2012; Bouckaert
373 et al., 2014). In this case, applying the transmission analysis separately to a sample of trees
374 from the phylogenetic posterior can help account for uncertainty (Didelot et al., 2014). However,
375 two problems remain: how to choose the tree prior in the phylogenetic reconstruction and how
376 to combine the results from the separate transmission analyses. A solution may be to see the
377 phylogenetic trees sampled in the first step as coming from a biased distribution, and correcting
378 for this using importance sampling in the second step, such that the separate transmission
379 analyses are correctly aggregated and the prior used in the first step is nullified (Meligkotsidou
380 and Fearnhead, 2007). On a more positive note, it should be noted that our two-step approach

381 has significant advantages both computationally and conceptually. Computationally, we were
382 able to analyse outbreaks with hundreds of cases in a matter of hours. Conceptually, working
383 with a fixed phylogeny allows us to explore much more complex models for transmission trees,
384 such as the partially sampled and ongoing scenarios. To date, no other transmission inference
385 approaches handle these difficult scenarios.

386 We have previously applied earlier versions of our approach to understanding a complex
387 tuberculosis outbreak in a largely homeless Canadian population (Didelot et al., 2014; Hatherell
388 et al., 2016), showing how reveals key individuals contributing to transmission and how its ability
389 to time infection events can be used to declare a waning tuberculosis outbreak truly over. Here,
390 we demonstrate our new methodology's ability to identify unsampled cases. Finding such cases
391 is critically important for tuberculosis control – not only does it allow us to seek out these
392 individuals and connect them with treatment, but it allows us to extend our case-funding efforts
393 to include a larger proportion of potentially exposed individuals. In our present analysis of the
394 Hamburg dataset, we found that the generation time was relatively rapid, with the majority
395 of infected individuals progressing to active disease and infecting others doing so within two
396 years, with many progressing to active disease almost immediately. This is important data for
397 outbreak management – if borne out by further reconstructions, it suggests a bound for the
398 time over which an individual who has been exposed to tuberculosis should be followed up.

399 In conclusion, we present a new method for the automated inference of person-to-person
400 disease transmission events from pathogen genomic data, one which accounts for the complex
401 and variable nature of sampling cases during an outbreak. When coupled to the routine
402 genomic surveillance of key pathogens now in place at many public health agencies, such
403 as Public Health England's new genomic approach to tuberculosis diagnosis and laboratory
404 characterisation (Pankhurst et al., 2016), our method has the potential to rapidly suggest the
405 contact network underlying an outbreak. Given the significant resources associated with a
406 contact investigation, any tool that can quickly assist in prioritising individuals for followup is
407 an important contribution to the public health domain.

408 **Acknowledgments**

409 This work was supported by the UK National Institute for Health Research Health Protection
410 Research Unit in Modelling Methodology at Imperial College London in partnership with Public
411 Health England (grant HPRU-2012-10080 to XD) and the UK Medical Research Council (grant
412 MR/N010760/1 to XD). JG holds a Canada Research Chair in Public Health Genomics and a
413 Michael Smith Foundation for Health Research Scholar Award. The funders had no role in study
414 design, data collection and interpretation, or the decision to submit the work for publication.

415 References

- 416 Anderson, R. M. and May, R. M., 1992. *Infectious diseases of humans: dynamics and control*,
417 volume 28. Wiley Online Library.
- 418 Azarian, T., Daum, R., Petty, L., Steinbeck, J., Yin, Z., Nolan, D., Boyle-Vavra, S., Hanage,
419 W., Salemi, M., and David, M., *et al.*, 2016. Intra-host evolution of methicillin-resistant
420 staphylococcus aureus usa300 among individuals with reoccurring skin and soft tissue
421 infections. *Journal of Infectious Diseases*, :jiw242.
- 422 Barry, C. E., Boshoff, H. I., Dartois, V., Dick, T., Ehrt, S., Flynn, J., Schnappinger, D.,
423 Wilkinson, R. J., and Young, D., 2009. The spectrum of latent tuberculosis: rethinking
424 the biology and intervention strategies. *Nat. Rev. Microbiol.*, **7**(12):845–55.
- 425 Becker, N., 1977. Estimation for discrete time branching processes with application to epidemics.
426 *Biometrics*, :515–522.
- 427 Biek, R., Pybus, O. G., Lloyd-Smith, J. O., and Didelot, X., 2015. Measurably evolving
428 pathogens in the genomic era. *Trends Ecol. Evol.*, **30**:306–313.
- 429 Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A.,
430 Rambaut, A., and Drummond, A. J., 2014. BEAST 2: a software platform for Bayesian
431 evolutionary analysis. *PLoS Comput. Biol.*, **10**(4):e1003537.
- 432 Cori, A., Ferguson, N. M., Fraser, C., and Cauchemez, S., 2013. A new framework and software
433 to estimate time-varying reproduction numbers during epidemics. *Am. J. Epidemiol.*,
434 **178**(9):1505–12.
- 435 Croucher, N. J. and Didelot, X., 2015. The application of genomics to tracing bacterial pathogen
436 transmission. *Curr. Opin. Microbiol.*, **23**:62–67.
- 437 Didelot, X., Gardy, J., and Colijn, C., 2014. Bayesian inference of infectious disease transmission
438 from whole genome sequence data. *Mol. Biol. Evol.*, **31**:1869–1879.
- 439 Didelot, X., Walker, A. S., Peto, T. E., Crook, D. W., and Wilson, D. J., 2016. Within-host
440 evolution of bacterial pathogens. *Nat. Rev. Microbiol.*, **14**:150–162.
- 441 Diel, R., Ru, S., and Niemann, S., 2004. Molecular Epidemiology of Tuberculosis among
442 Immigrants in Hamburg, Germany. *J. Clin. Microbiol.*, **42**(7):2952–2960.
- 443 Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., and Solomon, W., 2002. Estimating
444 mutation parameters, population history and genealogy simultaneously from temporally
445 spaced sequence data. *Genetics*, **161**(July):1307–1320.
- 446 Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A., 2012. Bayesian phylogenetics
447 with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.*, **29**(8):1969–1973.
- 448 Farrington, C. P., Kanaan, M. N., and Gay, N. J., 2003. Branching process models for
449 surveillance of infectious diseases controlled by mass vaccination. *Biostatistics*, **4**(2):279–95.
- 450 Fine, P. E. M., 2003. The Interval between Successive Cases of an Infectious Disease. *Am. J.*
451 *Epidemiol.*, **158**(11):1039–1047.

- 452 Fourment, M. and Holmes, E. C., 2014. Novel non-parametric models to estimate evolutionary
453 rates and divergence times from heterochronous sequence data. *BMC evolutionary biology*,
454 **14**(1):1.
- 455 Gilchrist, C. A., Turner, S. D., Riley, M. F., Petri, W. A., and Hewlett, E. L., 2015. Whole-
456 genome sequencing in outbreak analysis. *Clin. Microbiol. Rev.*, **28**(3):541–563.
- 457 Gire, S. K., Goba, A., Andersen, K. G., Sealfon, R. S., Park, D. J., Kanneh, L., Jalloh, S.,
458 Momoh, M., Fullah, M., Dudas, G., *et al.*, 2014. Genomic surveillance elucidates ebola
459 virus origin and transmission during the 2014 outbreak. *science*, **345**(6202):1369–1372.
- 460 Golubchik, T., Batty, E. M., Miller, R. R., Farr, H., Young, B. C., Lerner-Svensson, H.,
461 Fung, R., Godwin, H., Knox, K., Votintseva, A., *et al.*, 2013. Within-Host Evolution
462 of *Staphylococcus aureus* during Asymptomatic Carriage. *PLoS One*, **8**(5):e61319.
- 463 Grassly, N. C. and Fraser, C., 2008. Mathematical models of infectious disease transmission.
464 *Nat. Rev. Microbiol.*, **6**(6):477–87.
- 465 Green, P. J., 1995. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian
466 Model Determination. *Biometrika*, **82**(4):711–732.
- 467 Hall, M., Woolhouse, M., and Rambaut, A., 2015. Epidemic Reconstruction in a Phylogenetics
468 Framework: Transmission Trees as Partitions of the Node Set. *PLOS Comput. Biol.*,
469 **11**(12):e1004613.
- 470 Harris, S. R., Cartwright, E. J. P., Torok, M. E., Holden, M. T. G., Brown, N. M., Ogilvy-
471 Stuart, A. L., Ellington, M. J., Quail, M. A., Bentley, S. D., Parkhill, J., *et al.*, 2013.
472 Whole-genome sequencing for analysis of an outbreak of methicillin-resistant *Staphylococcus*
473 *aureus*: a descriptive study. *Lancet Infect. Dis.*, **13**(2):130–136.
- 474 Hatherell, H.-A., Didelot, X., Pollock, S. L., Tang, P., Crisan, A., Johnston, J. C., Colijn, C.,
475 and Gardy, J., 2016. Declaring a tuberculosis outbreak over with genomic epidemiology.
476 *Microb. Genomics*, **1**:10.1099/mgen.0.000060.
- 477 Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C., and Ferguson, N., 2014. Bayesian
478 Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data.
479 *PLoS Comput. Biol.*, **10**:e1003457.
- 480 Kingman, J., 1982. The coalescent. *Stoch. Process. their Appl.*, **13**(3):235–248.
- 481 Lloyd-Smith, J., Schreiber, S., Kopp, P., and Getz, W., 2005. Superspreading and the effect of
482 individual variation on disease emergence. *Nature*, **438**(November):355–9.
- 483 Meligkotsidou, L. and Fearnhead, P., 2007. Postprocessing of genealogical trees. *Genetics*,
484 **177**(1):347–358.
- 485 Mollentze, N., Nel, L. H., Townsend, S., le Roux, K., Hampson, K., Haydon, D. T., and
486 Soubeyrand, S., 2014. A Bayesian approach for inferring the dynamics of partially observed
487 endemic infectious diseases from space-time-genetic data. *Proc. R. Soc. B Biol. Sci.*,
488 **281**(1782):20133251.
- 489 Pankhurst, L. J., del Ojo Elias, C., Votintseva, A. A., Walker, T. M., Cole, K., Davies, J.,
490 Fermont, J. M., Gascoyne-Binzi, D. M., Kohl, T. A., Kong, C., *et al.*, 2016. Rapid,
491 comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: A
492 prospective study. *Lancet Respir. Med.*, **4**(1):49–58.

- 493 Paterson, G. K., Harrison, E. M., Murray, G. G. R., Welch, J. J., Warland, J. H., Holden,
494 M. T. G., Morgan, F. J. E., Ba, X., Koop, G., Harris, S. R., *et al.*, 2015. Capturing the
495 cloud of diversity reveals complexity and heterogeneity of MRSA carriage, infection and
496 transmission. *Nat. Commun.*, **6**:6560.
- 497 Roetzer, A., Diel, R., Kohl, T. A., Rückert, C., Nübel, U., Blom, J., Wirth, T., Jaenicke, S.,
498 Schuback, S., Rüsck-Gerdes, S., *et al.*, 2013. Whole Genome Sequencing versus Traditional
499 Genotyping for Investigation of a Mycobacterium tuberculosis Outbreak: A Longitudinal
500 Molecular Epidemiological Study. *PLoS Med.*, **10**(2):e1001387.
- 501 Romero-Severson, E., Skar, H., Bulla, I., Albert, J., and Leitner, T., 2014. Timing and order
502 of transmission events is not directly reflected in a pathogen phylogeny. *Mol. Biol. Evol.*,
503 **31**(9):2472–2482.
- 504 To, T.-H., Jung, M., Lycett, S., and Gascuel, O., 2016. Fast dating using least-squares criteria
505 and algorithms. *Syst. Biol.*, **65**(1):82–97.
- 506 Tong, S. Y. C., Holden, M. T. G., Nickerson, E. K., Cooper, B. S., Cori, A., Jombart, T.,
507 Cauchemez, S., Fraser, C., Wuthiekanun, V., Thaipadungpanit, J., *et al.*, 2015. Genome
508 sequencing defines phylogeny and spread of methicillin-resistant Staphylococcus aureus in
509 a high transmission setting. *Genome Res.*, **25**:111–118.
- 510 Wallinga, J. and Lipsitch, M., 2007. How generation intervals shape the relationship between
511 growth rates and reproductive numbers. *Proc. Biol. Sci.*, **274**(1609):599–604.
- 512 Worby, C. J., Lipsitch, M., and Hanage, W. P., 2014. Within-Host Bacterial Diversity Hinders
513 Accurate Reconstruction of Transmission Networks from Genomic Distance Data. *PLoS*
514 *Comput. Biol.*, **10**(3):e1003549.
- 515 Worby, C. J. and Read, T. D., 2015. 'SEEDY' (Simulation of Evolutionary and Epidemiological
516 Dynamics): An R package to follow accumulation of within-host mutation in pathogens.
517 *PLoS One*, **10**(6):1–14.
- 518 Young, B. C., Golubchik, T., Batty, E. M., Fung, R., Larner-svensson, H., Votintseva, A. A.,
519 Miller, R. R., Godwin, H., Knox, K., Everitt, R. G., *et al.*, 2012. Evolutionary dynamics
520 of Staphylococcus aureus during progression from carriage to disease. *Proc Natl Acad Sci*
521 *USA*, **109**:4550–4555.
- 522 Ypma, R., van Ballegooijen, W. M., and Wallinga, J., 2013. Relating Phylogenetic Trees to
523 Transmission Trees of Infectious Disease Outbreaks. *Genetics*, **195**:1055–1062.

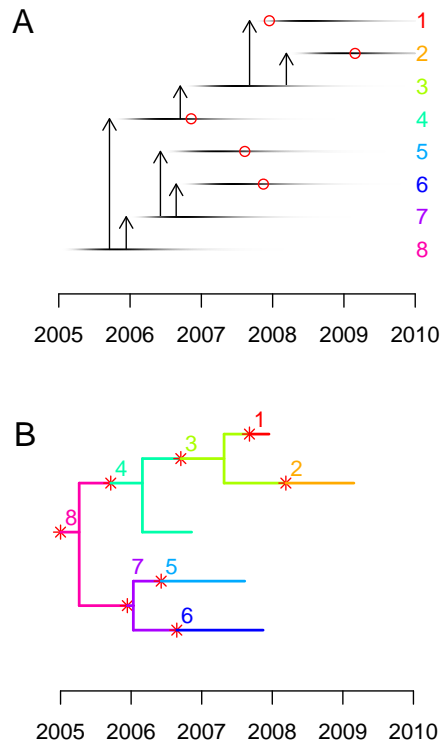


Figure 1. A: An illustrative example of transmission tree, with each horizontal line representing a case, and the darkness of each point representing their changing infectivity over time. Vertical arrows represent transmission from case to case. The red circles indicate which individuals were sampled (1, 2, 4, 5 and 6) and when. B: An example of colored phylogeny which corresponds to the transmission scenario shown in part A. Evolution within each host is shown in a unique color for each individual, as indicated by the labels and on the righthand side in part A. Red stars represent transmission events and correspond to the arrows shown in part A. Tips of the phylogeny represent sampled cases as shown by the red circles in part A.

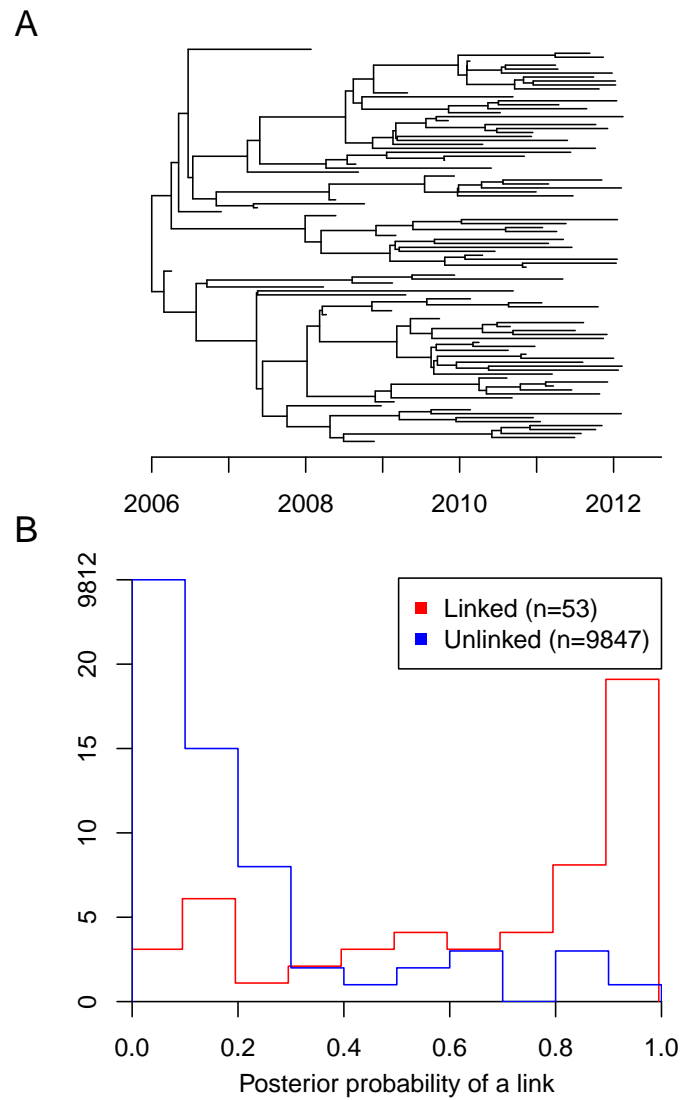


Figure 2. A: Timed phylogeny showing the relationship between 100 genomes sampled with density $\pi = 0.5$ in a simulated outbreak. B: Distribution of the posterior probability of direct transmission inferred by our algorithm for pairs of individuals in which a link existed in the simulation (red) and pairs of individuals which were not linked (blue).

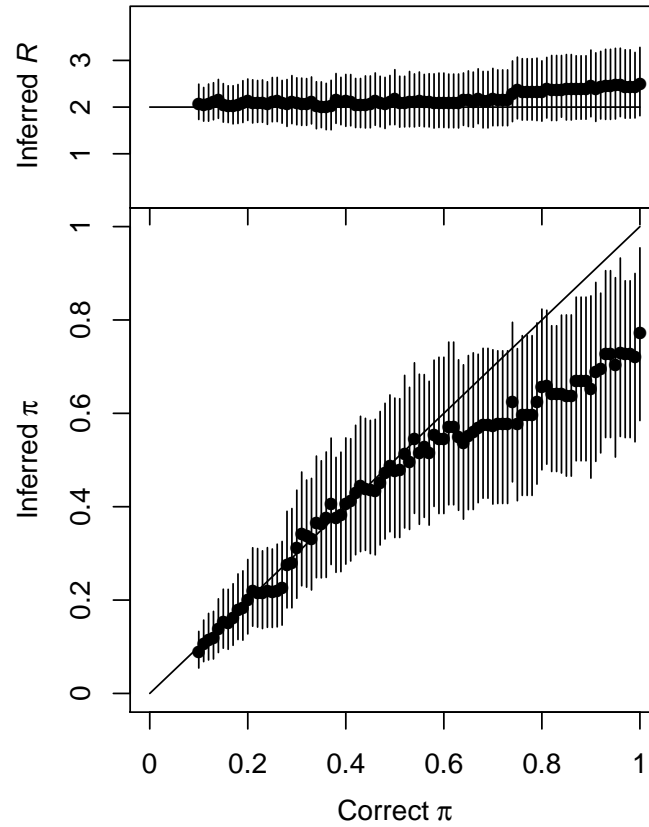


Figure 3. Inferred values of the reproduction number R (top) and the sampling proportion π (bottom) in simulated datasets for which the correct value of R is 2, and the correct value of π is increased from 0.1 to 1 (as shown on the x-axis). Dots represent the mean of the posterior sample and bars the 95% credibility intervals.

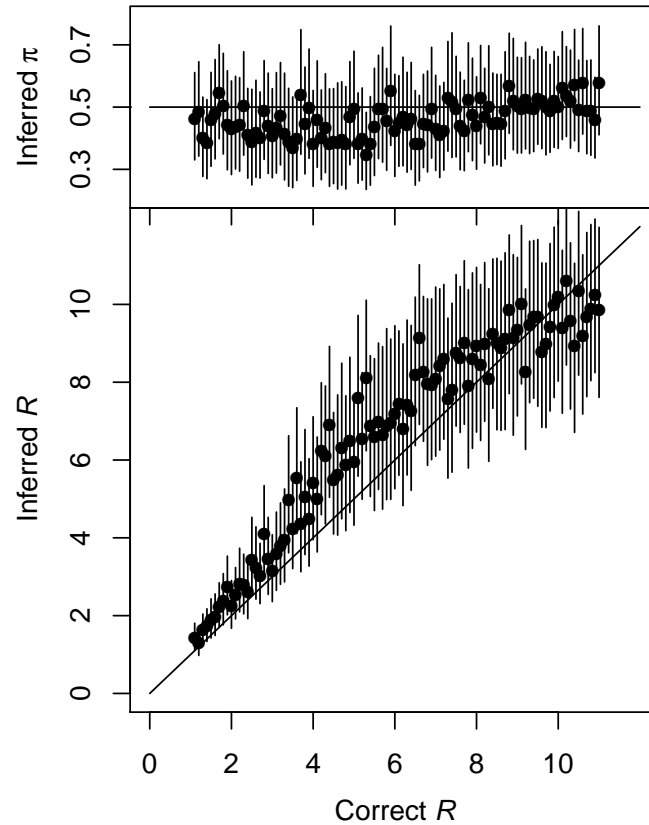


Figure 4. Inferred values of the sampling proportion π (top) and the reproduction number R (bottom) in simulated datasets for which the correct value of π is 0.5, and the correct value of R is increased from 1 to 11 (as shown on the x-axis). Dots represent the mean of the posterior sample and bars the 95% credibility intervals.

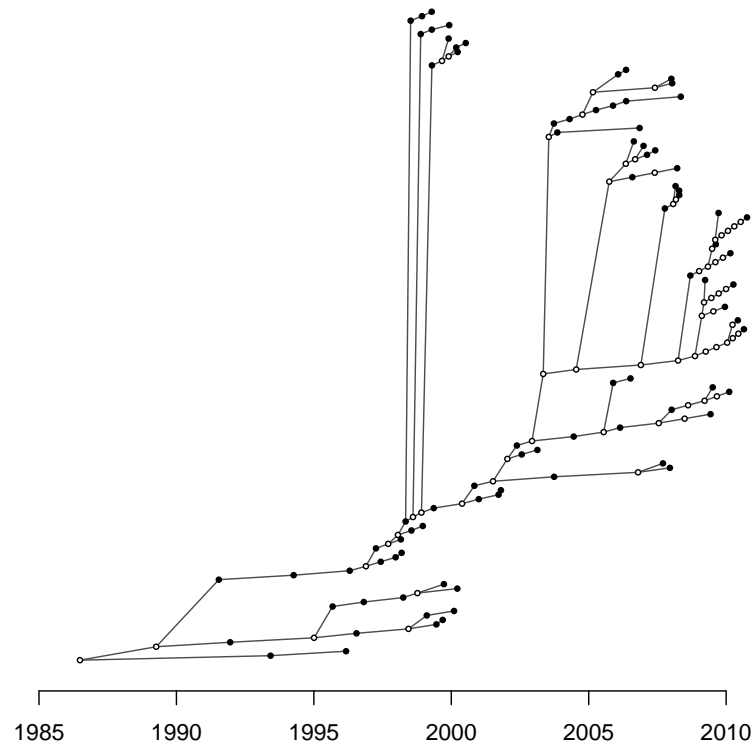


Figure 5. Consensus transmission tree for the tuberculosis outbreak. Filled dots represent sampled individuals and unfilled dots represent unsampled inferred individuals.

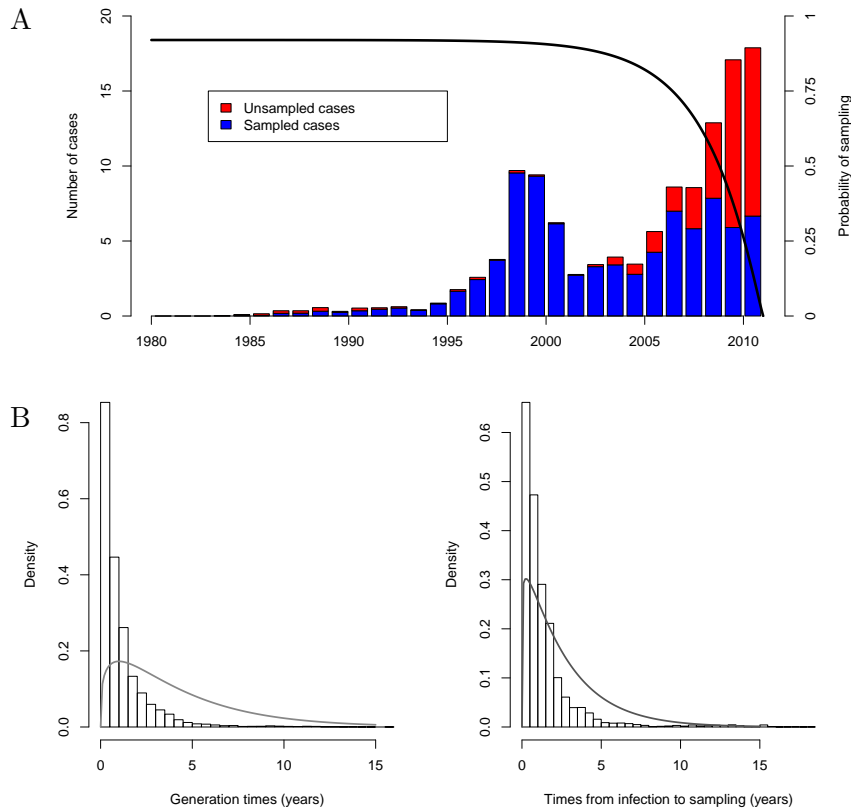


Figure 6. A: Outbreak plot showing the numbers of sampled and unsampled cases through time in the posterior sample of transmission trees. While the posterior estimate of π is 0.93, predicting that cases would eventually be detected with high probability, in the time period just before sampling ended, the inferred transmission trees contain a number of unsampled cases. The solid line represents the probability of sampling cases as a function of their infection time, given that observation stops at $T = 2011$. B: Posterior generation times and times between infection and sampling. Bars show histograms of the posterior quantities and solid lines show the related prior densities.